Estimating Net Urban Agglomeration Economies

With An Application to China[1]

Chun-Chung Au and J. Vernon Henderson

Brown University

August 2003

Abstract

This paper models and estimates net urban agglomeration economies for cities. Economic models of cities postulate an inverted-U shape of real income per worker against city employment, where the inverted-U shifts with industrial composition across the urban hierarchy of cities. This relationship has never been estimated, in part because of data requirements. China has the necessary data and context. We find that urban agglomeration benefits are high – real incomes per worker rise sharply with increases in city size from a low level. They level out nearer the peak, but then decline very slowly past the peak. We find that a large fraction of cities in China are undersized, due to strong migration restrictions and find large income losses from these restrictions.

1. INTRODUCTION

This paper develops a model for use in econometrically assessing net urban agglomeration economies, and then estimates the model with data on cities in China. This will be the first paper to assess net urban agglomeration economies; the China data and context have unusual features that make estimation possible. We will then use the results to explore costs of migration restrictions in China, which sharply curtail rural-urban migration and appear to leave most Chinese cities undersized.

[1] This work was funded in part by research grants from the World Bank and by the Population Studies and Training Center at Brown University. Part of the motivation for the paper arose from comments from Gilles Duranton.
JEL codes 00, P3, R0. Key words: urban agglomeration, scale economies, city size, China, migration.

Economic models of cities, where the number of cities in an economy is endogenous, postulate an inverted-U shape of real income per worker against city size, with the result that under appropriate institutions, cities will operate at or near the peak, where real income per worker is maximized (Henderson 1974, Helsley and Strange 1991, Black and Henderson 1999, Fujita, Krugman, Venables 1999, Duranton and Puga 2001). While there is an enormous literature examining, in particular, industry-specific scale externalities, and a smaller literature examining costs of specific types of urban diseconomies such as commuting and congestion costs (see Rosenthal and Strange 2004, and Moretti 2004 for reviews), no empirical paper has put the two together to estimate the net outcome, the inverted-U shape to real income worker.

Since we don't know what this inverted-U looks like, we don't know how quickly real incomes rise with agglomeration within a city, how quickly they diminish past the peak, and how the peak shifts across the urban hierarchy. Estimation will allow us to assess the extent of net agglomeration economies, and to assess, for example, the welfare costs of institutional or policy constraints and deficiencies that lead to over- or under-sized cities. The results will also have implications for policies that derive from notions of "optimal" city sizes.

There are two key reasons why net urban agglomeration economies have not been estimated to date. First, most countries like the USA do not collect GDP figures at the geographic level of an appropriately defined economic city, such as a metropolitan area. Second, theory suggests that, under free migration within a country, if particular cities are not at their peak, they will be to the right of the peak, due to either "stability" conditions in migration-labor markets or conditions on what constitutes a Nash equilibrium in migration decisions (Henderson and Becker 2001). With no cities to the left of the peak, it is then implausible to trace out empirically the shape of the inverted-U. China provides a data set and context that overcome these problems. China collects and publishes GDP figures at the level of the appropriately defined metro area, with a three-sector breakdown. Second, harsh migration restrictions sharply curtail in-migration to cities, so results indicate cities are spread all over the inverted-U , allowing us to identify its shape.

Given the appropriate data and context, four problems remain. In theory and practice there are many types of cities in an economy, where different types of cities produce different sets of products, have different production scale economies, and have different "efficient" sizes where output per worker is maximized. That is, there is not one inverted-U for cities, but many. We will develop a model that gives a way to characterize how the inverted-U shifts with industrial composition. The specification will be based upon patterns in the data and econometric results will support it.

Second, systems of cities models have no specific geography and cities no specific locations (except perhaps along a "featureless" line). Econometrically, we need to account for the effect of geography on inverted-U 's. In particular, cities in different locations have differential access to domestic and international markets and face different effective demands and prices. We incorporate into a system of cities model, the transport cost-varieties-monopolistic competition elements of the new economic geography (Fujita et al., 1999). Then we will be able to define how price varies with city supply and with city demand, or the market potential a city faces.

Third, estimation in any context of aggregate GDP-factor input relationships is plagued by endogeneity problems. LHS and RHS variables are mostly all endogenous. Traditional methods such as differencing to eliminate "fixed effects" and then instrumenting for endogeneity to contemporaneous shocks are plagued by problems. The error structure may poorly approximate fixed effects and past levels of covariates may be weak instruments for current changes, both in practice and in theory (e.g., Blundell and Bond, 1998). However the China context provides excellent instruments, for productivity relationships estimated in levels form. As we will detail, we estimate productivity relationships post-market reforms; but can instrument with certain pre-reform variables not driven by current types of unobservables, but determined under planning. Given accumulation processes in both migration and capital markets these will turn out to be strong, valid instruments.

The final problem is really a caveat. The model we develop has specific market institutions which may not be fully mimicked in China (or in the USA). Regardless of institutions, the variables in the meta-production function for a city are the same. Certain differences in institutions may affect the height but not the shape or peak point of the inverted-U's, while others may shift peaks. We will try to distinguish these, but any results ultimately must be interpreted for the institutions which apply to the data.

In section 2 of the paper we present the model, to be implemented econometrically. In section 3 we discuss the China context, data and econometric issues; and then we present results. In section 4, we conclude and illustrate the use of the results in evaluating the cost of China's migration restrictions within the urban sector.

2. THE MODEL

2.1 City Agglomeration

In this section, we present a simple model of productivity and industrial composition in a city. We will assume particular market institutions for derivations, but the intent is to develop a model that could be applied generally. We start with a derivation for an economy with just one type of city (and many cities of that type) and then generalize to n types.

Urban Production Technology

Cities produce final goods for sale to other cities (and potentially other countries) and intermediate service inputs which are non-traded, or sold only to local final good producers. All goods are varieties in the Dixit-Stiglitz-Eithier tradition, and sold under monopolistic competition. Final goods are shipped to other cities with iceberg-type transport costs.

In a representative city, the producer of final good variety $y(i)$ uses inputs of capital, k, effective labor, ℓ, and s_x varieties of intermediate input $x(i)$. This final good will be viewed as a "manufactured" product. Effective labor will be a critical concept, where the total effective labor of the city, L, will be less than the labor force, N, because of commuting costs described later. The producer faces a fixed cost, c, paid in units of the composite $y(i)$ Thus net output of the firm, $\tilde{y}$, is gross output, y, less c. Technology is

$$\begin{aligned} &\tilde{y} = y - c = A(\bullet)k^{\alpha}\ell^{\beta}\ \left(\int_{s_x} x(i)^{\rho}\,di\right)^{\gamma/\rho} - c \\ &\alpha + \beta + \gamma = 1, \quad 0 < \rho < 1 \end{aligned} \tag{1}$$

Producers have three sources of agglomeration economies. First will be the number of local varieties, s_x, of intermediate inputs which rise with city size. Note with symmetrical intermediate input producers y collapses to $A(\bullet)k^{\alpha}\ell^{\beta}x^{\gamma}s_x^{\gamma/\rho}$ where $\alpha + \beta + \gamma = 1$ but $\gamma / \rho > \gamma$. Second are local external scale economies, for which the typical parameterization is

$$A(\bullet) = AL^{\varepsilon} \tag{2}$$

where L is total effective city labor. Micro-foundations which aggregate up to a form like (2) include local information spillovers and search and matching economies, as reviewed in Duranton and Puga (2003). Finally the existence of transport costs of final output goods yield agglomeration benefits for consumers, as in the new economic geography.

Intermediate input producers in this model are viewed as producers of non-traded service inputs, used by final producers. Out-sourced business services are an obvious example; where, in China these are virtually completely non-traded across cities. And in the USA key out-sourced activities such as legal, accounting, finance, and insurance still are largely non-traded across metro areas (Schwartz 1993). To business services, one could add personal services and retail, which are also non-traded. Usually these are thought of as consumer final consumer final goods, and one can easily adjust the specification of preferences below, to incorporate this, with the same form to scale effects in the final aggregate meta-production function for the city. However we do not have the data to break out business from other services. Thus we keep things simple; and, perhaps, with a tip of the hat to Chinese history where state-owned enterprises typically provided most of these services to their workers, we leave consumption of all services in the production function – meals to feed the workers to work, so to speak. The producer of any variety faces a cost function defined in labor units of

$$\ell_x = f_x + c_x X \tag{3}$$

and sells his product in local monopolistically competitive markets. f_x is the fixed and c_x the marginal effective labor unit cost.

Demand for Final Output

Consumers nationally (or internationally) have preferences of the form

$$U = \left(\int y(i)^{\frac{\sigma_y - 1}{\sigma_y}} di \right)^{\frac{\sigma_y}{\sigma_{y-1}}} \qquad \sigma_y > 1 \tag{4}$$

Each producer in any city is a monopolistic competitor in national and international markets. Invoking standard results (see Overman, Redding and Venables 2001and Head and Mayer 2003 for reviews) the price, p_j, for a producer in city j is given by

$$p_j = MP_j^{1/\sigma_y} \; (y - c)^{-1/\sigma_y} \qquad , \tag{5}$$

where the price elasticity of demand, η_y, is $\eta_{\mathrm{y}} = -\sigma_y$. Market potential, MP_j, facing city j producers is

$$MP_j = \Sigma_v \; \frac{E_v I_v}{(1 + \tau_{jv})^{\sigma_{y-1}}} \tag{6}$$

where the sum is over all locations (markets) in the country (world). τ_{jv} is the iceberg cost of shipping a unit of output from j to v, E_v is total consumer expenditure in v, and I_v a price index. Assuming symmetry within cities, with preferences in (3),

$$I_v = \left[\Sigma_u \, s_y{}^u \; (p_u (1 + \tau_{vu}))^{1-\sigma_y} \right]^{-1}.$$

Again the sum is over all locations, $s_y{}^u$ is the number of varieties produced at location u and $p_u(1+\tau_{vu})$ is the effective price of varieties from location u in location *v*.

Firm Profit Maximization and Entry

In our representative city, a producer of any final output variety seeks to maximize profits

$$pA(\bullet)k^{\alpha}\ell^{\beta}(\int_{s_x} x(i)^{\rho}\, di)^{\gamma/p} - c - \int_{s_y} q(i)x(i)di - w\ell - rk.$$

w is the local wage rate; r is the fixed to the city cost of capital in national or international markets; and $q(i)$ is the local price of intermediate input variety $x(i)$. First order conditions on input choices, given $\eta_y = -\sigma_y$, with price defined in (5) are

$$MP^{1/\sigma_y}(y-c)^{-1/\sigma_y}\left(\frac{\sigma_y - 1}{\sigma_y}\right)\beta y/\ell = w \tag{7a}$$

$$MP^{1/\sigma_y}(y-c)^{-1/\sigma_y}\left(\frac{\sigma_y - 1}{\sigma_y}\right)\alpha y/k \ = r \tag{7b}$$

$$MP^{1/\sigma_y}\ (y-c)^{-1/\sigma_y}\left(\frac{\sigma_y - 1}{\sigma_y}\right)\gamma \mathrm{y}/(s_x x) \ = q \tag{7c}$$

The last condition (7c), while derived for a single variety, then anticipates intermediate input symmetry where y producers each purchase x of any variety and buy s_x varieties.

We assume entry into local final goods markets until profits are driven to zero at s_y producers/entrants in the city. Using (7a-7c) in the profit function set equal to zero yields the equilibrium output for a single y producer, where gross output is

$$y = \sigma_y \; c. \tag{8}$$

For intermediate good producers, the technology in (1) gives the derived demand benefits of usage of x in y. From standard Dixit-Stiglitz-Eithier results, x producers face a price elasticity, η_x, of $\eta_x \; = \; -(1-\rho)^{-1}$. Given the unit labor cost function in (3) and profit maximization, we assume entry of s_x producers such that profits are driven to zero. This implies standard expressions for, respectively, mark up of price, q, over marginal cost, per firm output X, and per firm labor usage where

$$q \; = \; \frac{wc_x}{\rho} \tag{9a}$$

$$X = \frac{f_x \rho}{(1-\rho)c_x} \tag{9b}$$

$$\ell_x \; = \; \frac{f_x}{1-\rho}. \tag{9c}$$

Labor Market Clearing

The two local markets are for labor and for intermediate inputs. Market clearing conditions are

$$s_x \ell_x + \; s_y \ell = L \tag{10a}$$

$$X \; = \; s_y x \tag{10b}$$

(10a) is a full employment equation given s_x producers of x in the city and s_y producers of the traded good. (10b) states that supply of any variety, X, equals demand, where s_y producers each buy x of the intermediate input.

Production Aggregates for the City

To solve for a reduced form expression for total net city output, which we will use in econometric implementation, we need to solve for certain aggregates, in particular, s_x and s_y. For reasons which will be apparent later we also want an expression for the ratio of value-added in manufacturing (y) to that in services (x).

To solve s_x and s_y first we need to solve for the use of ℓ by y producers. Into (7a), substitute (9a) for w, (7c) for q, (10b) for x, and (9b) for X to get

$$\ell \;=\; \frac{\beta}{\gamma}\,\frac{f_x}{1-\rho}\,s_x / s_y \tag{11}$$

Then from (10a) we can solve

$$s_x \;=\; \frac{\gamma}{\gamma+\beta}\,\frac{(1-\rho)}{f_x}\,L \tag{12}$$

To solve for s_y, into the production relationship from (1), we substitute for k from (7b), ℓ from (11), s_x from (12), x from (10b) and (9b), and y from (8). The result is

$$s_y \;=\; Q_0^{\frac{1}{1-\alpha}}\, MP^{\frac{\alpha/\sigma_y}{1-\alpha}}\, r^{-\frac{\alpha}{1-\alpha}}\, A(\bullet)^{\frac{1}{1-\alpha}}\, L^{\frac{\gamma/\rho+\beta}{1-\alpha}} \tag{13}$$

where Q_0 is a parameter cluster.[2]

For the city, the net output, or city output less borrowing costs is $(p\tilde{y} - rk)s_y$. Substituting for p and $\tilde{y}$ we have

[2] $Q_0 \;=\; \sigma_y^{-1}\,(\sigma_y - 1)^{\alpha(1-1/\sigma_y)}\; c^{\alpha(1-1/\sigma_y)-1}\; \alpha^{\alpha}\rho^{\gamma}c_x^{-\gamma}\gamma^{\gamma/\rho}\;\beta^{\beta}\;(\gamma+\beta)^{-(\beta+\gamma/\rho)}\;\;(f/(1-\rho))^{\gamma(1-1/\rho)}$

$$\text{net output} = (MP^{1/\sigma_y}\ (y-c)^{\frac{\sigma_y - 1}{\sigma_y}} - rk)\ s_y$$

By substituting in for rk from (7b) and for y from (8) net output becomes $MP^{1/\sigma_y}\ Q_1 s_y$ for Q_1 a parameter cluster.[3] From (13) then we have

$$\text{net output} = Q_2\ MP^{\frac{1}{\sigma_y(1-\alpha)}}\ r^{-\frac{\alpha}{1-\alpha}}\ A^{\frac{1}{1-\alpha}}\ L^{\frac{\gamma/\rho+\beta}{1-\alpha}} \tag{14}$$

defined in terms of units of effective labor, L. $Q_2 \equiv Q_0^{\frac{1}{1-\alpha}}\ \text{Q}_1$. Given the r which cities face, in section 4 on policy, we will examine the extent to which they are at a size which maximizes net output per worker.

In estimation what we observe in the data are city value-added and city capital stock. In our model total city value-added, VA, the dependent variable is given by $MP^{1/\sigma_y}\ s_y\ (y-c)^{\frac{\sigma_y - 1}{\sigma_y}}$. Since we observe K not r, we want this in primal form. We substitute into the expression for VA for s_y from (13). Then using (7b), $k = K/s_y$ and (13) again, we solve for r in terms of K and substitute for r into the revised VA expression. The result is[4]

$$VA = Q_3\ MP^{1/\sigma_y}\ A(\bullet)\ K^{\alpha}\ L^{\gamma/\rho+\beta} \tag{15}$$

Equation (15), once effective labor is defined, will become the basic equation, used to identify the parameters in (14).

Equation (15) includes the "meta-production" function $A(\bullet)\ K^{\alpha}\ L^{\gamma/\rho+\beta}$. If we used different market allocation rules, that affects the exact form of (15). For example, in our derivation the number of

[3] $Q_1 = (1-\alpha)\ ((\sigma_y - 1)c)^{\frac{\sigma_{y-1}}{\sigma_y}}$

[4] $Q_3 = Q_0\ \alpha^{-\alpha}\ (\text{c}\ (\sigma_y - 1))^{\frac{(1-\alpha)\ (\sigma_y - 1)}{\sigma_y}}$

input varieties is not optimal. Optimality could involve paying per firm fixed costs, $wf_{x,}$ from the local "public budget", something both China and the USA may approximate through subsidy programs. Under an optimal number of intermediate input varieties, (15) would look the same except Q_3 would change. In that case the institutional change only shifts the inverted-U up or down, with no impact on its shape and city size where the inverted-U is maximized. Less "innocent" changes in institutions (such as restrictions on s_x, or potentially, s_y) would involve more complex analyses and potentially more complex forms to (15). But in the end, the basic arguments would be MP, $A(\bullet)$, K, and L.

Finally for later reference we note the ratio of value-added in manufacturing to that in services. Value-added in the y sector is $\rho(y-c)\ s_y - qs_x x$ and in the x sector is $qs_x X$. Utilizing (7c), (5) and (8) we can show that

$$MS = (1-\gamma)/\gamma \tag{16}$$

where MS is the manufacturing to service ratio.

Effective Labor

So far there are only benefits from agglomeration in a city. To have disadvantages, the stylized model of that assumes commuting costs increase as they would in a monocentric city where everyone works in the Central Business District (CBD), which is surrounded by residents. If the CBD is a point, people live on lots of fixed size one, the city is circular (an equilibrium configuration, absent specific geography (e.g. a port)), and employment or population is N, then the radius of the city is $\pi^{1/2}$ $\mathrm{N}^{-1/2}$. People living at distance b from the city center spend t amount of working time to commute a unit distance (there and back), or face total commuting time costs of bt. Then total commuting time costs for the city are $\int_0^{\pi^{-1/2}N^{1/2}} 2\pi b\ (tb)\ db$ where $2\pi b\ db$ people live in the ring at distance b. Integrating we get total commuting time of $2/3\ \pi^{-1/2}tN^{3/2}$. Therefore for a population of N, effective labor for a city is[5]

$$L = N\ -(2/3\pi^{-1/2}\text{t})\,N^{3/2} \tag{17}$$

This parameterization is for a very specific configuration, and doesn't allow for congestion, multi-centered cities, and variable lot sizes. In general in all cases, total commuting resources rise with city size, even as cities move from being monocentric to multi-centered (Fujita and Ogawa, 1982). We use (17) in estimation.

City Objective Function

The objective function in section 4 we focus on is net output per actual worker, accounting for the translation of effective into actual labor. This is the real disposable income per worker in cities, after capital rents are paid. If an individual city is of "optimal" (2nd best given our market institutions) size for itself, it would want to maximize this magnitude, in a setting where there are many cities who compete for mobile workers in national labor markets. We will discuss this more below, accounting for the possibility that cities may be "differentiated" -- in different locations with different market potentials or different $A's$ in equation 2.

Utilizing (17), we specify real output per worker net of capital rents from (14a) now substituting in (2) where $A(\bullet) = AL^{\varepsilon}$. The result is

$$\text{net output per worker} = Q_2\ MP^{\frac{1}{\sigma_y(1-\alpha)}}\ r^{-\frac{\alpha}{1-\alpha}}\ A^{\frac{1}{1-\alpha}}\ (N - a_0N^{3/2})^{\frac{\varepsilon+\gamma/\rho+\beta}{1-\alpha}}\ N^{-1} \tag{18}$$

where $a_0 = 2/3\pi^{-1/2}t$. Eq. (18) is also the real wage per worker, accounting for worker rents and commuting costs.[6] Similarly from the estimating equation for value-added in (15)

[5] The formulation assumes land rents paid which vary inversely with commuting distances are collected and redistributed in the city. In China that is more explicit, since land rents are nominal. In either case the only resource cost is commuting time (and potentially out-of-pocket costs of commuting).

$$VA = Q_3\ MP^{1/\sigma_y}\ A\ K^{\alpha}\ (N - a_0 N^{3/2})^{\varepsilon+\beta+\gamma/\rho} \qquad (19)$$

Given estimates of (19), from (18) we an calculate the city size that maximizes net output per worker, or the size at the peak of the inverted $-U$. If we maximize the log of net output per worker with respect to N and set equal to zero, we get that at the peak

$$\overset{*}{N} = \left(\frac{\varepsilon+\gamma\ (1\text{-}\rho)/\rho}{a_0\ (\varepsilon+\gamma\ (1-\rho)/\rho + \ 1/2(\varepsilon+\beta+\gamma/\rho))}\right)^2 \qquad (20a)$$

Simple calculations show that (1) $\partial \overset{*}{N}/\partial\varepsilon\ >\ 0$ if $\beta > \varepsilon$, (2) $\partial \overset{*}{N}/\partial\rho < 0$, and (3) $\partial \overset{*}{N}/\partial\gamma > 0$ if $\beta\ (1-\rho) > \varepsilon\rho$. Not surprisingly as city scale externalities, ε, rise, "efficient" city size increases, (ruling out "super" scale economies where the externality exceeds labor's private return, in order to have multiple cities of determinant size). As the value of having more intermediate input varieties increases, or ρ declines, "efficient" city size increases. And as the role of intermediate inputs increases, or γ rises, "efficient" city sizes increase (again ruling out a form of "super"-scale economies). Note from (16), as γ rises and the city peak point shifts right, the manufacturing to service ratio declines. That fact leads into a discussion of the urban hierarchy, essential for econometric interpretation and implementation of (18), (19) and (20a).

2.2 The Urban Hierarchy and Econometric Implementation

We conceive of cities as being in an urban hierarchy with different types of cities, absolutely or relatively specialized in different types of traded good products. So there are textile cities producing textile varieties, steel cities producing steel product varieties, large market cities producing consumer

[6] Specifically, each worker with labor endowment of one earns $w(1-tb)$ + land rent income - $R(b)$, where $R(b)$ is the rent at location b where the worker lives. Rental income is total city rents $1/3\pi^{-1/2}\ t\ N^{3/2}$, split evenly among all residents. Given the equilibrium rent gradient, $R(b) =\ t(b_1 - b)$ where b_1 is the radius of the city, substitutions give (18) as the real wage.

services, and so on. A detailed description of such a hierarchy is in Black and Henderson (2003), with very detailed work in Alexandersson (1959) and Bergsman, Greenstone and Healy (1972).

To put this in a theoretical model with geography and market potential, there are two key components. First we need to re-specify preferences in equation (4) to be

$$U = \prod_{g} \left(\int y_g(i)^{\frac{\sigma_g - 1}{\sigma_g}} \, di \right)^{\frac{\mu_g \sigma_g}{\sigma_g - 1}} \tag{4a}$$

where each g is a different product, with many varieties of each product. It is common to assume σ_g is the same across products so only the consumption weights, μ_g, differ. Now the form to market potential (6) becomes more complicated if τ's vary by product; and one has to incorporate consumption weights into the I_v term. Econometrically, for China, we simply don't have sufficient data to initiate a complex form and we will approximately a version of (6).

The second component is to assume the $A(\bullet)$ function is either written as $AL_g^{\varepsilon_g}$ or just AL^{ε_g}. The first form says that scale externality benefits are product specific -- textile producers only learn from other textile producers (and their intermediate input suppliers). So ignoring transport costs, cities would be better off specializing to maximize scale benefits relative to urban commuting diseconomies. But even if just ε_g (or ρ_g) varies by product there would be different efficient city sizes for different types of products; and that alone would generate specialization. These are well known results in the literature (see Duranton and Puga, 2004 for a review). Transport costs of products across cities are a force encouraging industrial diversity (see Fujita, Krugman and Venables, 1999, Chapter 11). We do not propose to solve the general equilibrium problem of the degree of expected specialization for the specific geography in China nor do we have measures of specialization. However as noted, cities are relatively and even absolutely specialized in the urban hierarchy, although economic diversity increases with size. So the specialization tendencies have strong impacts in the modern world, relative to the agglomeration forces of transport cost considerations.

In the urban hierarchy in the literature, in a market economy with perfect migration, free capital markets, and developers and/or local government involved in formation of new cities, any city type g operates near its peak point to real output per worker, which is also the real wage. All cities face the same horizontal national supply curve of labor (as viewed by an individual city). As we move up the urban hierarchy, bigger cities have their peak points to net output per worker shifted right, peaking near the supply curve. In particular, with perfect divisibility of cities, many cities of each type, and all cities having identical amenities, A^i, each inverted-U for net output per worker is tangent to the supply curve at its peak point, as illustrated in Figure 1 later. If A^i's vary, say, then those with high A^i's operate to the right of their peak points in stable equilibria.

The issue is how to characterize this hierarchy in our data. For that we turn to a second fact about more modern systems of cities. As we move up the hierarchy the manufacturing to overall service ratio, MS, declines. In China the simple correlation coefficient between MS, and city employment is about -.20, based on the overall service sector which is dominated by retailing and personal services which tend to a fixed proportion of GDP across all cities. For the USA, Kolko (1999) details the patterns, separating business services from retail and personal services. For six city size categories the manufacturing to business service employment ratio declines monotonically from 2.95 at the bottom to .67 at the top size category.[7]

In terms of implementation, in estimating eq. (19), from (16), although we don't know γ_g, we know $\gamma_g = 1/(1+MS_g)$. So we could assume only γ_g changes across the urban hierarchy, rising as we

[7] As an illustration of this hierarchy, there are data on the spatial "product cycle", as reported in Fujita and Iishi (1994), on manufacturing electronics for the plants of big Japanese firms. Standardized production of generic TV sets occurs in small towns (perhaps outside of Japan) with little need for business service inputs. Production of semi-experimental products occurs in bigger cities and R&D and experimental production occurs in the largest metro area. Quite apart from the magnitude of information/knowledge externality issues, more experimental product requires more business services -- out-sourcing to programmers, designers, venture capital, advertising launching campaigns, etc.

move up the hierarchy (from (17)) with MS_g falling. We will estimate this exact form, so that in logs equation (19) becomes with rearrangement

$$\begin{aligned}\ln\ (VA) \ &= 1/\sigma_\gamma\ \ln(MP)\ +\ \ln A +\ \alpha\ \ln K \\ &+ (\varepsilon+\beta)\ \ln\ (N -\ a_o N^{3/2}) + \\ &1/\rho\ ((1+MS)^{-1}\ \ln\ (N - a_o N^{3/2}))\end{aligned} \qquad \text{(19a)}$$

Apart from the $\ln\ A$ term in (19a), the parameters of interest are $\sigma_y,\ \alpha,\ (\varepsilon+\beta),\ 1/\rho,$ and a_0. We will call (19a) the structural model.

However in (19), empirical results suggest ε_g also rises across the urban hierarchy, apart from changes in γ. Henderson (1988) relates estimated ε_g's for different industries to the average sizes of cities specialized in those products for Brazil, as well as the USA, finding a strong positive relationship. In Davis and Henderson (2003), scale elasticities for headquarter activities, found in larger cities, are found to be much larger than traditional estimates for manufacturing products found in smaller cities (see Rosenthal and Strange, 2004 for a review). Thus in (18) and (19), we presume that generally the cluster $\varepsilon+\beta+\gamma/\rho$ rises across the urban hierarchy as both ε and γ/ρ rise. Below we will parameterize that with a functional relationship $(\varepsilon+\beta+\gamma/\rho)\ =\ a_1\ -\ a_2\ MS\quad, a_2 < 0$ so that we estimate

$$\ln\text{ (net output)} = 1/\sigma_y\ \ln\ MP\ +\ \ln\ A + \alpha\ \ln K + (a_1 -\ a_2\ MS)\ \ln\ (N -\ a_0\ N^{3/2}). \qquad \text{(19b)}$$

In (19b) the presumption is that $a_1,\ a_2,\ > 0$ and $a_1 -\ a_2\ MS\ >\ 1-\alpha.$ We will call (19b) the semi-structural model. Given (19b), from (18) if we maximize net value per worker given the new parameterization, we get the peak for net value per worker. This is

$$\overset{*}{N} = \left(\frac{a_1 + \ \alpha\text{-}1\text{-}\ a_2 MS}{a_o(\alpha - 1 + 1.5\ (a_1 -\ a_2\ \ MS))} \right)^2 \tag{20b}$$

3. ESTIMATING THE INVERTED-*U*

We start with a brief description of the context: urban, economic and migration policies in China and the basics of the Chinese urban system. Then we discuss data and the variables appearing in (19a) and (19b). Then we turn to estimation issues and results.

3.1.1 Policy

Migration and Urban Policy

All migration in China is curtailed by the hukou system detailed in Chan (1994, 2000), (see also Au and Henderson, 2002). Under the system, you are "citizen" of the locality of which traditionally your mother is a citizen. Citizenship confers specific local benefits -- access to health care, free public education, legal housing, better access to jobs -- which non-citizens are not eligible for. To permanently migrate, you need to change citizenship. China authorizes about 18 million such changes a year, a number that has not changed for years and that involves a high proportion of urban-urban and rural-rural moves. For temporary migration, you can get a "visa" with varying degrees of hassle and substantial fees (Cai, 2000) to work in another location without local citizenship benefits there. Alternatively you can choose to migrate illegally and be subject to round-ups and deportation.

In our time period, focused on 1997, the estimated stock of temporary migrants outside their specific locality (legal or not) was estimated to be about 100 million, with only 60% of these away for longer than 6 months (Chan, 2000). But for moves (flows), only 32% were outside of the own-province and only 36% involved rural-to-urban moves. Despite popular perceptions, focused on the hard life of temporary migrants in large cities, migration seemed in 1997 to be sharply limited and mostly return, or round-trip migration. Real income rural-urban gaps are large, with over a threefold difference (Lin, Cai and Li, 1996).

China maintains this policy in part due to political pressure by urban residents, who fear vast influxes of peasants. But the policy is also consistent with long-term plans on urbanization, as reflected in the Sixth Five Year Plan (1981-85). That plan, which continues in part to guide urbanization, intended to sharply constrain growth of large cities, while permitting limited transfer of hukou from rural areas to towns of smaller cities. Evidence suggests that this planning combined with China's long-term aversion to large cities has distorted the Chinese size distribution of cities compared to other countries. Based on Henderson and Wang (2003), in 2000 China had only 9 metro areas with populations over 3 million, but another 125 with populations in the 1-3 million range, a ratio of numbers of cities in the two size categories of .072, compared to a worldwide ratio of .27. More generally, ranking cities by size from smallest to largest and calculating the cumulative share of urbanized population within a country, China's Gini of .43 is substantially less than that for the world (.56) and is smaller than all other individual large countries. Finally we note that planning in the early 1980's also thought in terms of a strict urban hierarchy, where the large ("sophisticated") lead the small. So, for example, only the largest coastal cities were initially to have access to new technologies and FDI, with technology then "trickling-down" the hierarchy. We will want to account for this in estimation.

Market Reforms

China from 1978 has undergone successive market reforms, as nicely summarized in Perkins (1994), for the period up to the early 90's. The reforms up to the early 1990's put agriculture and rural industrial production on a more free market basis. Our data cover the period 1990-1997, a period of rapid urban industrial reforms. These reforms removed the remaining props under state owned industry and exposed them to increasing competition. Mostly heavily hit were interior and northern heavy industry cities, especially under the reforms in 1993 continuing into 1994. These reforms moved most planning functions to a market basis and represent a break point in the data in terms of how outputs are evaluated. As part of the reforms, constraints on the service sector were removed, with the rapid growth in private sector services permitted. The result, in this very short period of time, is to dramatically shake up the urban system. In estimation, in terms of an instrumental variables strategy, we will utilize this '93-'94 split, viewing economic data in larger cities from 1990 as heavily influenced by planning and the history of planning, and economic data from 1997 as primarily driven by market forces.

3.1.2 The Urban System

We have data for 1990-1997 on about 225 prefecture level cities (including 4 "provincial level" cities).[8] These are the larger formal cities in China, for which a metropolitan area is well defined. Prefecture level cities govern large rural areas and in more extreme cases (such as provincial capitals) these may cover an area the size of the state of Connecticut. However, while data are given for the whole area (the "municipality"), they are also given separately for the urbanized portion, called the "city proper". The boundaries of the urbanized area are adjusted on an ongoing basis, to reflect urban expansion into rural areas.

Table 1 gives some basics on the 206 cities used in the estimating sample. From 1990-97, their populations grew on average by 2% a year, but their non-agricultural labor force grew by 3% a year. The differential reflects two things. In 1990, some city propers had agricultural populations that moved into non-agricultural employment in subsequent years. More critically, population numbers exclude both shorter-term immigrants and longer-term immigrants who work in the city but may live, for example, just beyond the boundaries of the urban area where they are able to find ("illegal") rural housing. Non-agricultural employment numbers may better capture urban expansion and we rely on them for our size measure.

[8] For some years from 1990-1996 we have data on county level towns, but we can't separate out rural and urban portions for these. The rural (agricultural) labor force accounts for half the population of these towns.

Table 1 shows that real output per worker grew at an incredible rate during the period; for prefecture cities, the average annual rate was about 6.5% a year. Finally over time the manufacturing to service ratio declines. The decline involves freeing of service activity around 1993-1994. In the data, over the 1990-93 period the ratio declines modestly in total by 4-5%; between 1993 and 1994 it declines by 24% (in part due to some redefinition of manufacturing as service activity); and from 1994 on it declines by 4-5% a year. As restrictions on private service sector are removed it takes off. In estimation, we use 1997 data, in order to allow market forces the greatest opportunity to be fully operational, especially in the service sector.

3.2 Estimation

3.2.1 Data and Variables

A complete description of data sources, the data and variables is given in the Appendix. Here we note the highlights in terms of estimating eqs. (19a) and (19b). City output is value-added in the 2nd (manufacturing) and 3rd (services and trade) sectors. The manufacturing to service ratio is the ratio of value-added in the 2nd to 3rd sector, where we note that we have no way to separate out business services from personal services and trade. Labor force is the labor force in the 2nd and 3rd sectors. Capital stock is the capital stock in the city of all "independent accounting units", and covers in the mid-1990's the capital stock of the state-owned sector and about half of the urban collections and private firms. We assume this captures virtually the entire productive capital stock. However we did experiment extensively with controls for the ratio of output of independent accounting to other units. These controls are insignificant in our results, and have no affect on other results. Here we simply use the capital stock with no controls.

In terms of other covariates, there remain the arguments in A and for $\ln MP$. For A we are looking for items that would affect the city-specific level of technology and labor force quality. We know the ratio in 1990 of adults with high school ("senior middle school") education; and use that as a control for the 1997 labor force quality. For city-specific technology, other than influences of labor force quality, we know accumulated (since 1990) real FDI by city (in US dollars). We use the ratio of accumulated FDI divided by labor force, to control for effective technology. That specification, as opposed to simply total FDI (or FDI per unit of capital stock), produced the most "stable" results -- a coefficient on FDI that didn't fluctuate with the details of the rest of the specification. It is consistent with the idea that technology transfer is not a "pure public good" at the city level, but diffuses (is congested) with city scale.

Finally for $\ln\ (MP)$, we need to construct a measure of market potential. We have no complete set of observations on I_v in eq. (6) and drop it. However official price induces show very little price fluctuation across localities in China. E_v is total local GDP in all 223 1997 prefecture level cities plus about 430 other cities (and their rural populations); the coverage is almost the entire economy. For distance discounting we approximate $(1+\tau_{jv})^{\sigma_y - 1}$ by d_{jv} where d_{jv} is the distance in 100's of miles from the center of locality j to that of v. If $\sigma_y = 8$ which is a "typical" cited value in the monopolistic competition literature, then at 100 miles the implied τ_{jv} is .09; at 1000 miles it is .35; and at 2000 it is .46. Given China's very poor internal transport this seems reasonable, as well as consistent with the literature (Overman, Redding and Venables, 2001, Davis and Weinstein, 2003, and Head and Meyer, 2004).

The next issue is how to deal with international income, E_R, where the discount factor is distance from city j to the China coast. For that we decompose $\ln\ (MP)$ into

$$\begin{aligned} \ln\ (MP_j) &= \ln\left(\sum_{v\in\text{China}} \frac{E_v}{(1+d_{jv})} + \frac{E_R}{(1+\ d_{j\text{ coast}})}\right) \\ &\approx \ln\ (MP_{j\text{ domestic}}) + E_R\ \frac{1}{MP_{j\text{ domestic}}\ (1+\ d_{j\text{ coast}})} \end{aligned} \qquad (21)$$

where

$$MP_{j\text{ domestic}} \equiv \sum_{v\in\text{China}} \frac{E_v}{(1+d_{jv})}.$$

Own locality E_j has $d_{jv} = 0.$ In (19), $\ln\ (MP_{j\ \text{domestic}})$ in theory has a coefficient of $1/\sigma_y,$ while the second term will have a coefficient of $E_R/\sigma_y.$

3.2.2 Econometric Issues

In specifying eqs. (19) for estimation, the missing component is the error structure, e_{jt}. In general in a market context, there are unobserved variables that affect productivity and hence input choices -- capital, and labor which migrates to the city in response to the fact that productivity shocks will affect wages offered in competitive local labor markets. The shocks we have in mind relate to unobserved labor force quality, local infrastructure relevant to producers, efficiency of governance, and local "business climate and leadership". Many of these may be fairly persistent in a market context and could be summarized in an error structure where $e_{jt} = ge_{jt-1} + \zeta_{jt}$, with g indicating the degree of persistence.

As will become apparent below, our big concern in choosing instruments will be endogeneity of local scale, or the labor force size. We articulate a specific process that generates labor force size and then discuss a strategy for instrumenting. In China, as discussed earlier, migration to cities is very costly, subject to fees and regulations and lack of urban housing and formal sector jobs for non-permanent residents. Most migration is also local, from, in particular, the rural parts of a municipality into its city proper.

To model this, assume each city offers a utility level to a resident which is a function of its real wage and local quality of life, Q_t^j, where Q_t^j, are consumer amenities potentially distinct from producer amenities, A_t^i. Real wages are related to the city's allocation of labor and capital and scale (dis)economies, determining labor marginal products in eq. (1). Thus city utility that can be offered to migrants is some function $U_t^j = U\ (K_t^j,\ N_t^j,\ A_t^j,\ Q_t^j).$ We expect $U_K,\ U_A,\ U_Q > 0$ but the sign of U_N is ambiguous (given the inverted $U-$shape to real wages). Focusing on the determinants of city population, N_t^j, we turn to the rural sector in the rest of the prefecture.

Utility in the rural sector is given by $R_t^j, = (K_t^{Rj}, \bar{N}_t^j - N_t^j, A_t^{Rj}, Q_t^{Rj})$. $\bar{N}_t^j$ is the population of the whole region of j, encompassing the urban and rural sectors, so the rural sector population is this total less the urban population, N_t^j. K_t^{Rj}, A_t^{Rj}, and Q_t^{Rj} are respectively the capital, production amenities (e.g., soil quality, rainfall) and quality life measures in the rural sector. $\partial R_t^j / \partial N_t^j$ may be greater than zero, implying diminishing real wages in the rural sector so R_t^j rises with out migration. If there were no migration restrictions N_t^j would adjust to equalize R_t^j and U_t^j. In China, we presume $U_t^j > R_t^j$ a differential sustained by migration restrictions, which operate as frictions that have the per person cost of in-migration rising as the rate of net in-migration to the city, $\dot{N}_t^j \equiv (dN_t^j / dt) / \dot{N}_t^j$, rises. As was common in China, local migration of the labor force not involve a residence change, but just very long commuting from the rural area to the city proper. At any instant the gap between urban and rural utility equals the cost of migration $m(\bullet)$, or

$$U_t^j - R_t^j = m(\dot{N}_t^j),\ m, m' > 0 \tag{22}$$

Of course, absent naïve expectation, U_t^j and R_t^j, should represent the present values of utility in the urban versus rural sectors. This equation (with the LHS in present value terms) is the specification of migration frictions in the USA (in part due to rising costs of housing with in-migration in the short run), used in Mueser and Graves (1995) and Rappaport (2000). Eq. (22) provides (1) a link between amenities, Q_t^j, A_t^j, and Q_t^{Rj}, and local employment and (2) a link between past levels of amenities and of migration and current total local employment, through migration accumulations.

The framework sets the stage for identification. We assume historical (1990) variables such as the capital to labor ratio, FDI, observed amenities, and industrial composition were determined by planning.[9] 1990 is well before the major urban industrial reforms of 1993-94. Specifically these variables were not influenced by unobservables affecting market productivity today; planning decisions were based on ideological and political criteria. However migration even in 1990, especially within the prefecture, did respond to these unobservables affecting current and anticipated wages. Some of these historical unobservables may persist in influence to 1997 (i.e., in $e_{jt} = ge_{jt-1} + \zeta_{jt-1}$, is not small). Thus 1990 local labor force and quality of labor force (education) measures as well as scale of output are not considered as candidates for instruments. Instruments are 1990 amenity, capital intensity, industrial composition, and geographic variables,[10] determined by planning not market forces. These 1990 variables are strong instruments, because of accumulation processes (first stage F's averaging about 55 with a minimum over 17). The size of the current labor force is strongly related to past conditions and accumulated migration, given (22). Similarly past capital stock strongly influences the current stock, given similar accumulation processes.

We tested our presumptions in two ways. First informally we looked at post 1995 residuals and 1990 instruments finding very low correlations. Second we conducted Sargan tests focused on an original linear version of the model (see eq. (23) below). These pass readily, providing (only if) we omit 1990 scale variables -- population, labor force, or GDP, as well as 1990 education. We estimate eq. (19a) and (19b) by non-linear 2SLS for 1997, using the instrument list in fn. 9. In some specifications we experimented with including 1996 observations, but decided that we wanted to rely just on 1997 in order to allow market reforms more time to have their impact. Adding in 1995 observations led to serious

[9] In theory, ratio variables, K/N and MS, are not influenced by unobservables in A. That is, MS is given by $(1-\gamma)/\gamma$ and K/N $(= K/(s_y\ell + s_x\ell_x)$ is also just a function of parameters. Of course there are measurement error issues for 1997 values of these variables, that instrumenting hopefully takes care of.

[10] Instruments (all for 1990) are land area of the whole prefecture, books per capital, doctors per capita, telephones per capita, roads per capita, two regional dummy variables, the capital/labor ratio, the MS ratio, area time the MS ratio, dummy for FDI or not, FDI/labor force, ratio of sales of independent accounting to all units, ratio of agricultural VA of the whole municipality to ratio of city proper non-agricultural VA, and the two variables in (21). For the last, for instruments, own locality E_j is omitted in the summations -- it is a "toxic" local scale variable.

deterioration in Sargan test statistics in the linear model, indicating perhaps that "planning observables" persisting from 1990 may have still been affecting the calculation of local GDP and input choices.

3.3 Results

Basic results are given in Table 2, with an interpretation of scale effects in Table 3. Results in Table 2 are the IV results, with OLS results in the Appendix in Table C. The main effect of IV estimation is on scale variable coefficients, our main concern from the earlier description of migration responses.

In order to interpret all results we start by noting the results on capital intensity. In Table 2, the coefficient, α, takes values .44 and .46. While these may seem high, they are consistent with results based on micro data on Chinese technology, with a history of Soviet style capital-intensive planned production (see Jefferson and Singhe, 1999). Given this rock solid coefficient, we first focus on scale-urban hierarchy results and then on results on market potential and the arguments of A.

3.3.1 Scale and Hierarchy Effects

The Structural Model. The two columns in Table 2 represent results on the two basic models: -- eqs. (19a) and (19b). We start with the structural model, where only γ varies as we move across the urban hierarchy, according to eq. (16). While the results are okay, they are not supportive of this specific model.

First and most critically in (19a) the coefficient on the last term is $1/\rho$. The estimate is .387, but $0 < \rho < 1$, so the result contradicts the model. Second, as a way of explanation, we note that, given $\alpha = .46$, $\varepsilon + \beta = .59$, and $\beta = 1 - \gamma - \alpha$, the results imply $\varepsilon - \gamma = .05$. The data (eq. (16)) imply a γ of about .4 which would imply a scale effect, ε, of .45 which is off the chart. Now imposing eq. (16) is problematic. The γ that varies across cities relates to business services, while personal services and trade are usually a fixed fraction of GDP. With the right data, we could distinguish two types of services. Personal and trade would have a fixed γ_p in either the production function ("feeding the workers" is part of production in the city), or, equivalently in terms of the reduced form meta-function approach, the utility function, while business services would have a variable γ_b. Here we simply can't make that distinction. So we employ a more flexible "semi-structural" model in column (2), and set aside the results in column (1).

<u>The Semi-Structural Model</u>. The semi-structural model yields very plausible coefficients. We discuss these before turning to an interpretation of the exact scale effects for this model. First in column (2), based on (19b), $\varepsilon + \beta + \gamma / \rho = a_1$ - a_2 MS = .787 $-.0503$ MS. Given $\alpha + \beta + \gamma = 1$ and $\alpha =$ .44, this implies $\varepsilon + \gamma\ (1-\rho)/\rho = .15$ for an average value of MS of 1.45 (see Appendix). If

$\gamma = .4$ (consistent with $MS = 1.45$) and $\rho = .8$ (consistent with an elasticity of substitution in production in eq. (1) of $\sigma = 5$, from Davis and Henderson, 2003) that implies an ε of .06, which is most plausible for a typical city. Second, the coefficient on $N^{1.5}$ which is a_0 implies total commuting costs, $2/3\pi^{-1/2}t\ N^{3/2}$, are .03 $N^{3/2}$. For a typical city with a labor force of 500,000 and a population of 1.0 million, that implies (for N = 70 in 10,000's), 10.6 of the labor force of 50 is used up in commuting activity – about 20%. Again this seems reasonable in a developing country.

These coefficients allow us to calculate in eq. (20b) the peak point where net output per worker in (18) is maximized. These results are given in Table 3, part a which shows how these peak points shift as the manufacturing-to-service ratio varies. The table shows the nice decline in city employment where net output per worker is maximized, as the manufacturing to service ratio rises. The largest most service intensive cities ($MS = .6$) have peaks at an employment of 1.3m or population of 2.6m. This seems small, given the sizes of modern metro areas. But few Chinese cities are in this range. While the stated objective of larger Chinese cities is to have $MS < 1$, less than 24% of prefecture level cities met that objective in 1997; and only 6 cities have MS values less than .6, a value that we might think of as being more typical in a market economy for a large city.. The MS ratio has a mean of 1.4, and 18% of cities have values in excess of 2.0. Chinese cities remain very manufacturing intensive with MS values ranging up to 4 or more. At high MS values, the employment values for the peak point tail off sharply.

Table 3 also shows 95% confidence intervals on the employment size for the peak, based on applying the delta method. The confidence bands are quite wide, which given the nature of the exercise is not surprising. Still as we will see in section 4, many cities will fall outside these wide confidence intervals, generally being undersized. Actual city sizes lie both to the left and right of the point estimates of where the peaks lie. Of the 206 cities, 179 are to the left of their peak, and 27 are to the right. Finally we note we tried many variants of (19b), including altering the exponent of N from 1.5 (or trying unsuccessfully to identify a different exponent to 1.5) and different forms to how $\varepsilon + \beta + \gamma / \rho$ varies with MS. Eq. (19b) worked better than anything else we tried.

Net Urban Agglomeration Economies

In Table 3 part b and in Figure 1, we illustrate the magnitudes of net urban scale economies, for a low and high value of MS . Table 3 gives percent variations in net output, or real wage, per worker from eq. (18), for MS =1; while Figure 1 plots absolute values of real wage against city employment size for MS =1 and MS =2.7 for real wage normalized to 18,000 yuan per year at the peaks to the inverted-U. Several things are apparent. First there are enormous agglomeration economies. Moving from a city of 100,000 to 1.17m for MS =1 raises real-output per worker by 40%, and much more if one starts at a lower size such as 50,000. Second, most agglomeration benefits are realized by a size that is, say, half the size at the peak. Moving from 590,000 to 1.17m only increases real output per worker by 5%. Third, agglomeration benefits in small types of cities (MS =2.7) accumulate very rapidly compared to larger types of cities (MS =1).

Fourth, the effect of being oversized is small, compared to being under-sized. For MS =1, from a peak size of 1.17m if one subtracts 880,000 people, real output per worker falls by 16%; but, if one adds 880,000, it only falls by 6%. To get a 16% loss on the right side of the peak, one would need to increase employment size by 150% of its size at the peak. Real output per worker has a long flat portion near the peak, and real output per worker drops very slowly past the peak. This has implications for free market analysis of city sizes, with differing amenities across cities. Cities with better market potential or amenities will have their inverted-U 's shifted up. With free migration in Figure 1 and, say, a horizontal supply curve of people at 18000 yuan to any city, then the typical city with MS =1 peaks at 18,000. For MS =1, a special city with high amenities will have a peak above 18000 at the same size (1.17m). With free migration, that city's size will be at the point past its peak where its real output per worker intersects the horizontal supply curve at 18000. Given how slowly real output per worker declines past the peak, this could be a very large size.

Table 3 also has implications for any notions of "optimal city size". For any MS , first there are large error bands on the size where real output per worker peaks, so there is no precision in setting optimal city size. Second, being off the mark, by, say, 50% is not highly costly. Third, what is "optimal" in a world of heterogenous urban amenities for most cities is to be to the right of their peak points. While the "typical" city might be near its peak, an efficient allocation would have cities with better amenities operating to the right of their peaks.

Robustness of Scale Effects

To check on these effects, we looked at how VA per worker varied in a non-structural model, with scale-hierarchy variables. First we plotted a graph of VA/N against N. This shows, perhaps, an overall modest inverted-U but it is basically flat. In a market context with free migration and competitive city formation in national land development markets, we would expect to find a flat line. As discussed above, with different kinds of cities, each type would operate near the peak point for that type, to offer roughly the same real wage, as in Figure 1. So a typical city of each type would have an inverted-U that peaks near a horizontal line, representing the going national real wage clearing national labor markets. China does not have free migration. But there is no particular reason to expect a specific shape. As should be clear by now to find inverted-U shapes to VA/N as a function of city scale, we need to control for city type by controlling for industrial composition.

To do that we simply generalize and combine how parameters differ across city types and the expression for effective labor in eq. (17). In effect we estimate a generalized meta-production relationship and focus on "plotting" VA per worker against N vs. MS. Based on extensive experimentation, we report on the form

$$\begin{aligned} &\ln\ (VA \text{ net output per worker}) = 1/\sigma_y\ \ln\ MP + \ln\ A + \alpha\ \ln\ (K/N) \\ &\qquad\qquad a_3\ N - a_4\ N^2 - a_5\ N \bullet MS \end{aligned} \tag{23}$$

In (23) a term $N^2\ MS$ has a zero coefficient in estimation and we don't include it. The presumption is that $a_{3,}\ a_4,\ a_5 > 0$, and $a_3\ - a_5\ MS > 0$), so

$$\dot{N} = \left(\frac{a_3\ - a_5\ MS}{2a_4} \right) \tag{24}$$

For this relationship, to first examine it informally, we combined 1996-1997 to increase sample size and broke the sample into septiles based on MS values. We then did OLS regressions of the model in (23), dropping the $MS * N$ term, to calculate $\overset{*}{N}$ for each MS interval. For the lowest MS septile, $\overset{*}{N}$ is at 2.3m workers. At the second it jumps to 4.3m, but then after it declines monotonically taking values respectively of 2.4m, 1.4m, 1.3m, .60m and .28m. Then we formally estimated eq. (23), including the $MS * N$ interaction for the 1997 sample. This is a linear model which we estimate using the same instruments as in fn. 9, by standard GMM methods. Results on coefficients of common variables are similar to those in Table 2. For the scale hierarchy portion, coefficients (and standard errors) are $a_3 = .0110$ (.00299), $a_4 = .0000156$ (.00000602), and $a_5\ = .00473$ (.00130). Peak points for $MS\ = .6$, 1.0, 1.4, 1.7, and 2.0 are (in 1000's) 2618, 2013, 1408, 954, and 50. By MS 2.3 there is no peak. The peak points around typical values of MS of 1.5 to 1.7 are similar to the column (2) Table 2 model, as reported in Table 3. However now as we move MS up or down from there, the shifts in the peaks are much more dramatic. In particular as MS falls, the peak shifts much more to the left so that at $MS = .6$

the peak is at 2.6 million compared to half of that for the structural model. However, for this general curve fitting, a problem is that beyond MS =2.3, which covers about 9% of the sample, there is no inverted-U in terms of point estimates, a problem the structural models don't have. Despite these differences the portion of cities to the left and right of their estimated peak points is similar in the structural versus curve fitting models.

3.3.2 Other Results

For other results, we return to Table 2, column (2). We start with market potential. The coefficient on domestic MP is an estimate of $1/\sigma_y$, and the implied $\sigma_y = 1.71,$ which is the elasticity of demand for a city's product. The coefficient seems high and the elasticity low, but it is not implausible. In terms of effects a one-standard deviation increase in $\ln\ (MP_{\text{domestic}})$ increases VA by 14%, a big effect. For the corresponding variable for international market potential, an increase of one-standard deviation in that variable increases VA by 10%. However the estimates suggest international market potential itself is small. For a city, average domestic market potential is about 13500 units. For the international variable, the coefficient equals E_R / σ_y, but E_R is normalized (multiplied by 1000). Accounting for the normalization and given $\sigma_y = 1.71,$ $E_R = 5624$ and international market potential $E_R\ (1+\tau_{j\ \text{coast}})^{-1}$ for the average $\tau_{j\ \text{coast}}$ city is 803. Note this "international market potential" reflects the fact that the estimated E_R is the real E_R further discounted by transport costs from China's coast to international markets, as well as trade barriers. China's domestic markets are its key element, for any city.

For technology variables, there is education and accumulated FDI per worker. For the latter, a one-standard deviation increase leads to a 11% increase in VA, a very large effect. Technology transfer through FDI seems to bring high productivity benefits, perhaps a reason why the Chinese subsidize these transfers. Cities favored by policy in the early 1990's as FDI targets gained a significant advantage. Finally there is education. Unfortunately we do not have 1997 values and have to rely on 1990 values.

The coefficient is positive but not significant. A one-standard deviation increase in education based on the point estimate increases VA by 4%. Interactions of these variables with scale variables produced insignificant, small effects.

Finally we note we tried a variety of other potential amenity measures. Once the market potential variables are accounted for, distance to the coast itself has no effect. And details such as distance to a major highway or navigable river have no effects. Kilometers of paved road per person in a city has a significant positive coefficient in non-IV estimation but an insignificant (negative) coefficient in IV estimation, a fairly standard result for public infrastructure. One interpretation of the IV result is that a zero coefficient means public infrastructure is generally at an optimal level, where slight increases or decreases then have no effect on productivity.

4. Under-Sized Cities

In the paper, we have estimated the inverted-U shape function of real output per worker against city scale, allowing the inverted-U to shift with city type, or industrial composition. Moving from very small relative size cities to appropriately sized ones for a given industrial composition, results in enormous productivity gains. However large upward deviations in size beyond the peak result in modest productivity losses.

The results have policy implications for China and we turn to these now. The results in Table 2 can be used to calculate a peak point for each prefecture city in China where net output per worker is maximized. And for each we can calculate a 95% confidence interval. In 1997, 80 cities had employment sizes outside the 95% confidence interval. Seven were significantly over-sized, which reflects the low efficient sizes predicted for the small sample of large cities with low MS values, in the semi-structural model. But the other 73 cities are significantly under-sized. The curve fitting exercise in eqs. (23) and (24) finds even more significantly under-sized Chinese cities – 139 in all.

In Table 4, we show a summary of the calculated percent loses in net output per worker, for the sample from operating at a size away from the point estimate of the peak. From eq. (18), we calculate

$$\ln\ (\text{net } \overset{*}{\text{output}}/\,N)\ \text{-}\ \ln\ (\text{net output}/\,N)$$
$$= \frac{1}{1-\hat{\alpha}}\left\{(\hat{\alpha}-1)\ (\ln\ \overset{*}{N} -\ \ln\ N) + (\hat{a}_1 - \hat{a}_2)\ MS\left[\ln\ (\overset{*}{N}\ -\ \hat{a}_0\ \overset{*}{N}{}^{1.5})\ -\ \ln\ (N\ -\ \hat{a}_0\ N^{1.5})\right]\right\}$$

where an asterisk is the value at the peak for N. The ratio MS is held constant.

Table 4 shows welfare losses for cities ranked by percentile losses. Although all but 8 cities operate at a size that is more than 10% from peak size, for 50% of cities welfare losses are fairly small, as we would expect from Table 3 and Figure 1. At the 50th percentile, that city's loss is 9.5%, in terms of *VA* per worker. Overall the average loss is 12.1%. However for 25% of cities, we are talking about losses over 18%, and for 10% of cities, losses over 26%. There are 6 under-size cities with losses over 40%. Allowing migration to these cities would allow them to operate much more efficiently. But that of course is only the tip of the iceberg. The gains to the migrants relative to their current wages would be enormous.

There are many caveats on this exercise. Foremost is that in a free migration, constrained efficient world, not all cities would operate at their peaks. Those with better natural amenities and market access would have larger sizes, ceteris paribus, drawing in workers and firms to operate to the right of the peak. Solving out how heterogeneous urban sites would be allocated across different types of cities in a context of real geography is a theoretical exercise yet to be attempted. But in a huge county like China, with an essentially uncountable number of viable urban sites, it is unclear how much natural amenity differentials across urban sites really matter in talking about under-sized cities. What is clear is that free migration would result in large increases in city sizes and productivity gains, as well as ultimately more cities.

References

Alexandersonn, G. (1959), "The Industrial Structure of American Cities", Lincoln: University of Nebraska Press.

Au, C.C. and J.V. Henderson (2002), "How Migration Restrictions Limit Agglomeration and Productivity in China", NBER Working Paper #8707.

Bergsman, J., P. Greenston, and R. Healy (1972), "The Agglomeration Process in Urban Growth", Urban Studies.

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

Black, D. and V. Henderson (2003), "Urban Evolution in the USA," Journal of Economic Geography (in press).

Blundell, R. and S. Bond (1998), "GMM Estimation With Persistent Panel Data", IFS Working Paper No. W99/4.

Cai, Fang (2000), Zongguo Liudong Renkou Wenti (The Mobile Population Problem in China), Henan People's Publishing House: Zhengzhou.

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

Chan, K.W. (2000), "Internal Migration in China: Trends, Determination, and Scenarios", University of Washington, report prepared for World Bank (April).

Davis, J. and J.V. Henderson (2003) "Headquarters' Location Decisions", Brown University mimeo, http://www.econ.brown.edu/faculty/henderson/papers/Agglomeration_of_Headquarters51303.pdf.

Davis, D. and D. Weinstein (2003), "Market Access, Economic Geography and Comparative Advantage: An Empirical Assessment", Journal of International Economics, 59, 1-23.

Duranton, G. and D. Puga (2001), "Nursery Cities", American Economic Review, 91, 1454-1463.

Duranton, G. and D. Puga (2004), Micro-Foundations of Urban Agglomeration Economies, in J. V. Henderson and J.-F. Thisse (eds.) Handbook of Regional and Urban Economics, Volume 4. Amsterdam: North-Holland, http://dpuga.economics.utoronto.ca/papers/urbanagg.pdf.

Fujita, M., P. Krugman and A.J. Venables (1999), The Spatial Economy, MIT Press.

Fujita, M. and R. Ishii (1999), "Global Location Behavior and Organizational Dynamics of Japanese Electronics Firms", in A.D. Chandler et al. (eds.) The Dynamic Firm, Oxford University Press, 344-383.

Head, K. and T. Mayer (2004), "The Empirics of Agglomeration and Trade", Handbook of Regional and Urban Economics, Vol. 4, J.V. Henderson and J-F Thisse (eds.) North Holland, forthcoming, http://www.econ.brown.edu/faculty/henderson/neat1.pdf.

Helsley, R. and W. Strange (1990), "Matching and Agglomeration Economies in a System of Cities", Journal of Urban Economics, 20, 189-212.

Henderson, J.V. (1974), "The Size and Types of Cities," American Economic Review.

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

Henderson, J. V. and H.G. Wang (2003), "Urbanization and City Growth," Brown University mimeo, http://www.econ.brown.edu/faculty/henderson/papers/Urbanization_and_City_Growth0803.pdf.

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

Kolko, J. (1999), "Can I Get Some Service Here: Information Technology, Service Industries, And the Future of Cities", Harvard University mimeo.

Lin, J.Y., F. Cai and Z. Li (1996), The China Miracle: Development Strategy and Economic Reform, The Hong Kong Centre for Economic Research and The International Center for Economic Growth, The Chinese University Press.

Moretti, E. (2004). "Human Capital Externalities in Cities", Handbook of Urban and Regional Economics, Vol. 4, J.V. Henderson and J-F Thisse (eds.), North Holland, forthcoming, http://www.econ.brown.edu/faculty/henderson/cities5.pdf.

Mueser, P. and P. Graves (1995), "Examining the Role of Economic Opportunity and Amenities In Explaining Population Redistribution", Journal of Urban Economics, 37, 176-200.

Overman, H.G., S. Redding and A.J. Venables (2001), "The Economic Geography of Trade, Production, and Income: A Survey of Empirics," LSE Handbook of International Trade, J. Harrington and K. Choi (eds.) Basil Blackwell, forthcoming.

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

Rosenthal, S. and W. Strange (2004), "Evidence on the Nature and Sources of Agglomeration Economies", Handbook of Urban and Regional Economics, Vol. 4, J.V. Henderson and J-F Thisse (eds.), North Holland, forthcoming, http://www.econ.brown.edu/faculty/henderson/WillAndStuart.pdf.

Schwartz, A. (1993), "Subservient Suburbia: The Reliance of Large Suburban Companies on Central City Firms for Financial and Professional Services", Journal of American Planning Association, 59(3), 288-305.

Table 1. Prefecture Level Cities

	1990	1997	Growth
Average population of the city proper (1000's)	922	1087	18%
Non-agricultural employment (1000's)	415	527	27%
Value-added per worker in non-agricultural sector (1990 yuan)	6389	10588	66%
Manufacturing to service (VA) ratio	2.17	1.44	-51%

Table 2. Results for Urban Productivity
(standard errors in parentheses)

		“semi
	“structural” model	structural” model
a for capital	.458**	.441**
	(.0882)	(.0816)
$(\varepsilon + \beta)$	.593**	
	(.176)	
$1/\rho$	.387**	
	(.184)	
$-a_0\ (= 2/3\ \pi^{-1/2}t)$	-.0350**	-.0300**
	(.00702)	(.0109)
a_1		.787**
		(.130)
a_2 (for MS)		-.0503**
		(.0241)
$1/\sigma_y$ (ln MP_{domestic})	.629**	.584**
	(.128)	(.119)
% h.s. education	.00247	.00521
	(.00565)	(.00490)
FDI per worker	.0684**	.0723**
	(.0307)	(.0300)
constant	-1.51	-.781
	(1.82)	(1.61)
N	206	206
R^2	.917	.921

** significant at 5%; * significant at 10% level.

Table 3. Urban Agglomeration

(a) City Employment at the Peak to

Net Output Per Worker

MS	**.6**	**1.0**	**1.4**	**1.7**	**2.0**	**2.5**	**3.0**	**3.7**
peak point in 1000's	1307	1174	1034	924	811	616	418	163
lower* 95% confidence interval	536	448	264	68				
upper 95% confidence interval	2079	1907	1803	1781	1787	1832	1828	1580

* A blank indicates a negative lower bound.

(b) Agglomeration Benefits

$(MS = 1)$

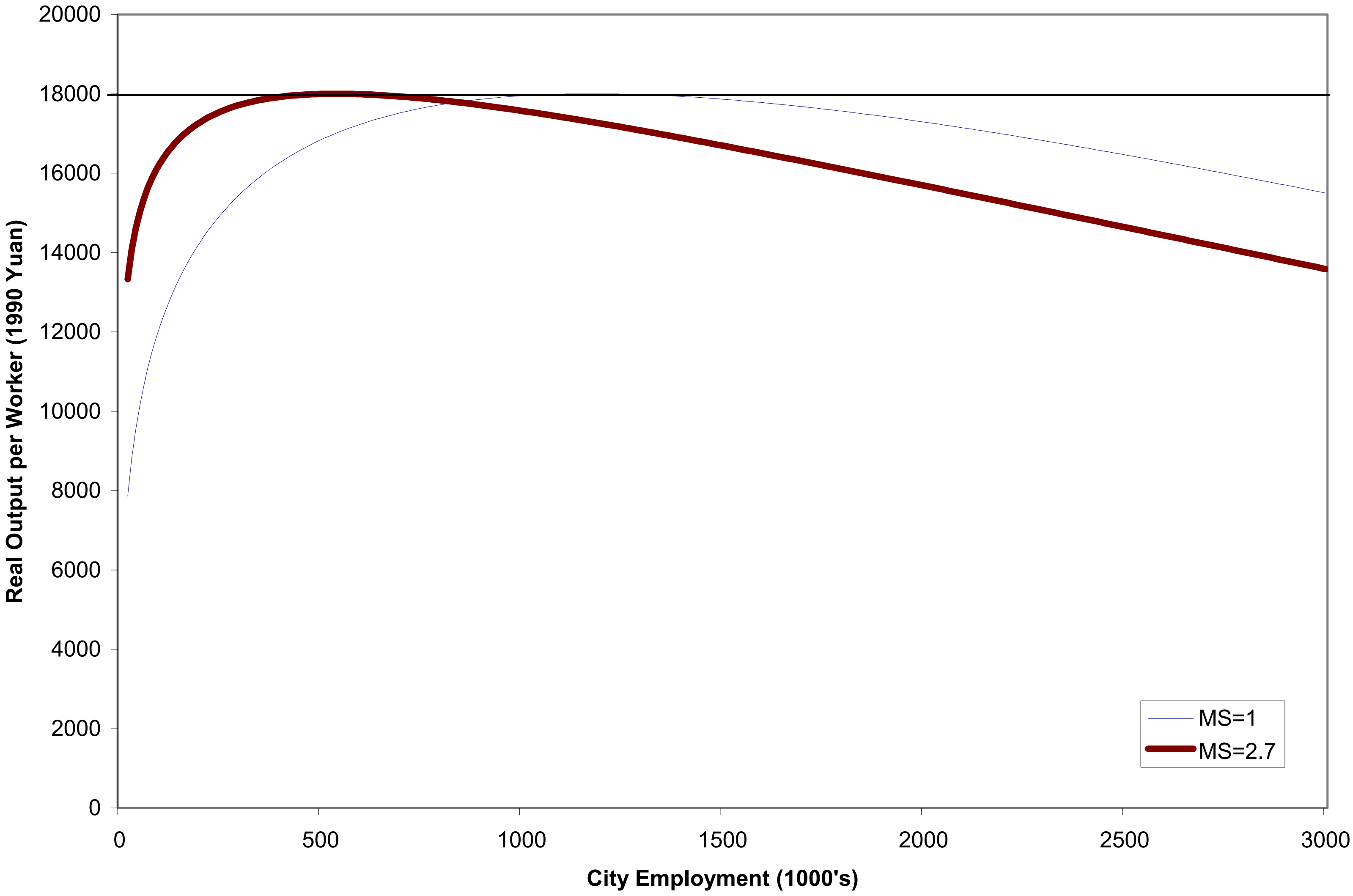
Employment 1000's	**20**	**50**	**100**	**290**	**590**	**880**	**1,170**	**1,470**	**1,760**	**2,060**	**3,000**
Decline (%) in net output per worker away from the peak at $\overset{*}{N} = 1{,}170$	-83%	-57%	-40%	-16%	-5%	-2%	0	-1%	-3%	-6%	-17%
Decline /%) (increase) (%) in size from peak at $\overset{*}{N} = 1{,}170$	-98%	-96%	-91%	-75%	-50%	-25%	0	+26%	+50%	+76%	+156%

Table 4. Percent Losses in Net Output per Worker from

Operating Away From the Peak

Percentiles of Cities (ranked by loss):	
first	largest loss (%)
5% of cities	0.12
10%	0.42
25	2.8
50	9.5
75	18
90	26
95	37
100	80

Figure 1. The Inverted-U for Cities



Appendix

The Urban Sector

City level data used in our analysis come from several sources. Most economic and amenity variables were taken from the 1991 to 1998 annual volumes (for data years 1990 to 1997) of the *Urban Statistical Yearbook of China* (hereafter *Yearbook*)[1], and *Cities China 1949-1998*. The latter includes a compilation of selected data in 1990 to 1997 for prefecture level cities from the *Yearbook* volumes and a complete history of new city establishment and changes in administrative area of all cities during the period. Distance proxies are measured with a ruler from *Map of China* in units of approximately 100 miles. Highway access is read directly from the same map (occasionally with help from a more detailed map). Educational attainment is aggregated from the *China County-Level Data on Population (Census) and Agriculture, Keyed to 1:1M GIS Map, 1990.* GDP figures are not available for 1992 and 1993 from the *Yearbook*, but are documented for prefecture level cities in *Cities China 1949-1998*. It should be noted that all city level data that we use are those of the more confined city proper (shi qu) rather than the municipal district (di qu). The city proper corresponds to an "urbanized area" in the USA, or the urbanized portion of a metropolitan statistical area. In our analysis of prefecture level cities, we have excluded three oil-dominant cities[2], and a minimal set of outlying city-years[3] based on extraordinary year-to-year change in output per capita, capital-labor ratio or manufacturing to service ratio which are likely the results of misdocumentation. For 1997, we then start with a base sample of 218 out of 223 prefecture level cities. The estimating sample is 206, where missing cities have missing data, or more particularly can't be linked to 1990 for instrumental variable construction. Brief descriptions of the variables used in our analysis are in Table A.

Three issues should be noted here. First, capital is original book value of capital of industrial enterprises with independent accounting systems. In attempts to control for the share of independent accounting units among production units in a city, we would want to use the share of value-added of independent accounting units in the city economy. The data needed to derive this share is available only in 1997. Second, for comparison of real growth of output (GDP), we use the provincial level urban resident consumer price index to deflate nominal GDP's. The index is taken from the Price Indices section of the annual *China Statistical Yearbook* in the relevant period. To compare the real output across cities, we have to assume comparability based on nominal prices in a certain year (1990 in our case).

[1] A combined volume was published for 1993 and 1994.

[2] Daqing, Dongying and Karamay.

[3] Jining of Shandong province and Songyuan in 1997.

Table A. Description of Variables

Variable	Description	Source(s)
population	- population at the end of the year	S1, S2
output of a city	- GDP of city in 2nd and 3rd sectors at current prices	S1, S2
manufacturing to service ratio (MS)	- ratio of GDP in 2nd sector to GDP in 3rd sector	S1, S2
employment	- number of persons employed in 2nd and 3rd sectors	S1, S2
capital	- original value of capital of industrial enterprises with independent accounting system	S1
output (value-added) of independent accounting units	- gross industrial output value (value-added of industry) of industrial enterprises with independent accounting system at current prices[4]	S1
FDI	- accumulated sum of foreign direct investment (foreign capital actually used) since 1990	S1, S2
roads per capita	- paved area of all roads with width greater than 3.5 meters	S1, S2
% high school	- percentage of population aged 6+ that has completed senior middle school or above	
distance to coast	- shortest horizontal distance from coast, measured in centimeters from map S4	S3, S4
distance to provincial capital	- horizontal distance from capital of province in which a city is located, measured in centimeters from map S4	S3, S4
on highway	- dummy for cities with access to highway (the highest category of all roads on map)	S3, S4
area (1990)	- built-up area in city proper	S2
doctors per capita (1990)	- number of medical doctors per capita	S2
books per capita (1990)	- number of books in public library per capita	S2
telephone per 100 persons	- number of telephones per 100 persons	S2
ratio of municipal agriculture to city value-added	- ratio of total GDP in 1st sector in municipal area to total non-agricultural GDP in city proper	S1, S2

Sources of Data

S1. State Statistical Bureau, Urban Social and Economic Survey Team [Guojia Tongjiju Chengshi Shehui Jingji Diaocha Zongdui], *Urban Statistical Yearbook of China* [*Zhongguo Chengshi Tongji Nianjian*], Beijing: China Statistics Press, 1991 to 1998 (annual volumes).

S2. State Statistical Bureau, Urban Social and Economic Survey Team [Guojia Tongjiju Chengshi Shehui Jingji Diaocha Zongdui], *Cities China 1949-1998* [*Xin Zhongguo Chengshi Wushi Nian*], Beijing: Xinhua Press, 1998.

S3. *Map of China* [*Zhongguo Quantu*], Haerbin Map Press, 3rd ed., February 1999. #1280529-158

S4. *Transportation Map of China* [*Zhongguo Jiaotong Yingyun Licheng Tuji*], Beijing: People's Communication Press, 2000. ISBN 7-114-03553-5.

S5. State Statistical Bureau, *China Statistical Yearbook* [*Zhongguo Tongji Nianjian*], Beijing: China Statistical Publishing House, 1996, 1998 and 1999 (annual volumes) and other relevant years.

[4] Calculated from industrial output value realized per 100 yuan of fixed assets at book value (value-added realized per 100 yuan of fixed Assets at book value) and fixed assets at book value of industrial enterprises with independent accounting system

Table B. Urban Variable Means and Standard Deviations

	All Prefecture Level Cities in 1997		
	mean	standard deviation	
output per worker	23079	11077	
capital per worker	30335	18197	
employment (1000's)	510	661	
total value sales/sales of independent acct. units 1996	1.42	.522	
% high school	21.9	8.48	
distance to coast	6.09	6.06	
on highway (highest road category)	.66	.47	
manufacturing to service ratio (GDP)	1.46	.739	
accumulated FDI since 1990 ($)/capital (yuan)	.0374	.0685	1

Table C.

	structural model	**“semi-structural” model**
α (ln (capital)*	.425**	.429**
	(.0437)	(.0425)
$(\varepsilon + \beta)$	.548**	
	(.0972)	
$1/\rho$	.128*	
	(.0761)	
$-a_0$ $(N^{1.5})$	-.00525	-.0109
	(.0438)	(.0528)
a_1		.616**
		(.103)
$-a_2$ MS		-.0215**
		(.0108)
$1/\sigma_y$ (ln (MP_{domestic}))	.533**	.521**
	(.0813)	(.0793)
E_R/σ_y	3.24**	3.31**
	(1.241)	(1.23)
% high school education	.00442	.00464*
	(.00308)	(.00298)
FDI per worker	.0720**	.0712**
	(.0176)	(.0178)
constant	.143	.240
	(.030)	(.880)
N	206	206
(adj) R^2	.924	.925