Review of Economic Studies (2006) **73**, 549–576 0034-6527/06/00210549$02.00

Are Chinese Cities Too Small?

CHUN-CHUNG AU

and

J. VERNON HENDERSON

Brown University

First version received January 2004; *final version accepted November* 2005 (*Eds.*)

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 and then decline very slowly past the peak. We find that a large fraction of cities in China are undersized due to nationally imposed, strong migration restrictions, resulting in large income losses.

1. INTRODUCTION

This paper develops a model for estimating net urban agglomeration economies that drive the existence of urban agglomerations and are the key force in urbanization in developing countries. The paper estimates the model with data on cities in China. This is the first paper to econometrically assess net urban agglomeration economies; the Chinese data and context have unusual features, which make estimation possible. We then use the results to explore the costs of migration restrictions in China, which sharply curtail migration and appear to leave many Chinese cities significantly undersized. The framework developed could be applied, correspondingly, to assess the issue of (presumably) oversized cities in countries with very different institutions, such as Thailand, Egypt, or Indonesia.

Economic models with an endogenous number of cities postulate an inverted U shape of real income per worker against city size (Henderson, 1974; Helsley and Strange, 1990; Black and Henderson, 1999; Fujita, Krugman and Venables, 1999; Duranton and Puga, 2001). While there is an enormous literature examining industry-specific scale externalities, which foster urban agglomeration, and a smaller literature examining costs of specific types of urban diseconomies, which limit city sizes (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 per worker.

The paper develops a model of the key components concerning scale economies and diseconomies internal to a city, as well as incorporating the effects of inter-city trade costs following the new economic geography. A key issue concerns how to specify an estimating model that accounts for the fact that there is an urban hierarchy in a country, with more than one type of city. We will report estimates of a structural model, although we will focus on the results of non-structural estimation to the shape of the inverted U in doing our assessment that Chinese cities are too small.

Once we know the shape of the inverted U, we can determine 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. We can then start to assess the welfare costs of institutional or policy constraints and deficiencies that lead to oversized or undersized cities. The results also have implications for policies discussed in the informal literature on the empirics of optimal city size (*e.g.* Tolley, Gardner and Graves, 1979). That literature does not tackle the problem head on, to provide an assessment of both net urban agglomeration economies and the optimality of sizes of the various types of cities in an economy.

There are two key reasons why net urban agglomeration economies have not been estimated to date. First, most countries like the U.S. do not collect and report gross domestic product (GDP) figures at the geographic level of an appropriately defined economic city, such as a metropolitan area. Countries, such as Brazil that report such numbers impute them from state-level GDP data, where the state-level numbers already reflect other imputations. 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–labour markets or conditions on what constitutes a Nash equilibrium in migration decisions (Duranton and Puga, 2004). With no cities to the left of the peak, while one could still in theory estimate the components of urban-scale effects and then use these to extrapolate the whole shape of the inverted U, results might seem less than convincing when trying to infer the shape of the curve to the left of the peak.

China provides a data-set and context that overcomes these problems. Local statistical bureaus have, for years, collected data on all enterprises in their local area and report GDP figures at the level of the appropriately defined metro area, with a three-sector breakdown. While doubts are often expressed as to the quality of national data in China, which may reflect politicized aggregations of data submitted by local and other statistical bureaus, local data are of high quality as discussed further below. Second, harsh migration restrictions sharply curtail in-migration to cities, so results indicate cities are spread all over the inverted U, allowing us to better identify its shape and then ultimately to argue that Chinese cities generally are undersized.

Given the appropriate data and context, four problems remain. First, in theory and in 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 sizes where output per worker is maximized. That is, there is not one inverted U for cities, but many. The structural model will show how one can characterize directly with the data, the way in which the inverted U shifts with industrial composition.

Second, systems of cities models have no specific geography and cities no specific locations (except perhaps along a "featureless" line). In theory and empirically, we need to account for the effect of geography on inverted U's. 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), so as to define how prices vary 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. Typically, both L.H.S. and R.H.S. variables are 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 after the market reforms in the early and mid-1990's, which directly exposed the huge state-owned urban industrial sector to market competition and opened up much of the business service sector to private firms. However, we can instrument with particular pre-reform, planning variables, which are not affected by current types of unobservables affecting market outcomes. Given accumulation processes in both migration and capital markets, historical variables will turn out to be strong 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 U.S.). 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 the peak. We will try to distinguish these, but results ultimately must be interpreted for the institutions that 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 apply the results to examine 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 start with an economy with just one type of city (and many cities of that type) and then generalize in Section 2.2 to n types in an urban hierarchy.

2.1.1. Urban production technology. Cities produce final differentiated 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 and Stiglitz (1977) tradition, 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_y, effective labour, ℓ_y, and s_x varieties of intermediate input $x(i)$. As is appropriate in the case of China, this final good, which is traded across cities, is viewed as a "manufactured" product. Effective labour will be a critical concept, where the total effective labour of the city, L, will be less than the number of people in the labour force, N, because of the commuting time costs described later. The producer faces a fixed cost, c_y, paid in units of the composite $y(i)$, needed to determine the size and number of firms in the local market. Thus, net output of the firm, $\tilde{y}$, is gross output, y, less c_y. Technology is

$$\tilde{y} = y - c_y = A(\cdot)k_y^{\alpha}\ell_y^{\beta}\left(\int_{s_x} x(i)^{\rho}\,di\right)^{\gamma/\rho} - c_y \tag{1}$$

$$\alpha+\beta+\gamma = 1, \quad 0 < \rho < 1.$$

Producers have three sources of agglomeration economies. First are local external-scale economies. While we test alternative specifications, the typical parameterization in the literature, which we utilize is

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

In (2) L is total effective city labour. 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 (2004). The second source of scale effects is the number of local varieties, s_x, of intermediate inputs, which will rise with city size. Note that with symmetrical intermediate input producers, y collapses to $A(\cdot)k_y^{\alpha}\ell_y^{\beta}(xs_x)^{\gamma}s_x^{\gamma(1-\rho)/\rho}$, where $\alpha+\beta+\gamma = 1$. The term $s_x^{\gamma(1-\rho)/\rho}$

indicates scale effects from having more varieties of local intermediate inputs. Finally, the existence of transport costs of final output goods modelled below yields implicit 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. Outsourced business services are the obvious example where, in China, these are virtually completely non-traded across cities, and in the U.S. key outsourced activities such as legal, accounting, finance, and insurance still are largely non-traded across metropolitan areas (Schwartz, 1993). To business services, one could add non-traded labour-intensive production of local intermediate manufacturing inputs, such as special order parts and components. And then there are personal services and retail, which are also non-traded. Usually personal and retail services are thought of as final consumer goods, and one can easily adjust the specification of preferences below to incorporate these, 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 non-traded service variety faces a cost function defined in labour units of

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

and sells his or her product in local monopolistically competitive markets. f_x is the fixed, and c_x the marginal effective labour unit cost.

2.1.2. Demand for final output of a producer. To solve the model, we need to know the demand for labour and capital by producers in the city, a derived demand dependent on final demand for city products. This tells us how p_y, the price of a final good variety for a local monopolistic producer, varies with the producer's output. To model this, assume consumers nationally (or internationally) have preferences of the form

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

Each producer in any city is a monopolistic competitor in national and international markets. Utilizing standard results (see Overman, Redding and Venables, 2003; and Head and Mayer, 2004, for reviews) the price, $p_{y,j}$, for a producer in city j is given by

$$p_{y,j} = \text{MP}_j^{1/\sigma_y} \ (y - c_y)^{-1/\sigma_y}, \tag{5}$$

where the price elasticity of demand is $\eta_y = -\sigma_y$, which is used to assess derived demands of local producers for intermediate service inputs. Market potential, MP_j, facing city j producers is

$$\text{MP}_j = \sum_{\upsilon} \frac{E_\upsilon I_\upsilon}{\tau_{j\upsilon}^{\sigma_y - 1}}, \quad \text{where} \ \ I_\upsilon = \left[\sum_u s_{y,u} \ \ (p_{y,u}\tau_{\upsilon u})^{1-\sigma_y}\right]^{-1}, \tag{6}$$

where the sum is over all locations (markets) in the country (world). $\tau_{j\upsilon}$ is the iceberg cost factor of shipping a unit of output from j to υ, E_υ is total consumer expenditure in υ, and I_υ is a price index where all producers operate symmetrically within cities, given preferences in (4). In the price index, the sum is over all locations, $s_{y,u}$ is the number of varieties produced at location u, and $p_{y,u}\tau_{\upsilon u}$ is the effective price of varieties from location u in location υ. Later in the empirical section we will devote considerable attention to the empirical implementation of (6).

2.1.3. Effective labour. The final key piece of the model for a single city concerns the definition of effective labour. So far we have only benefits from agglomeration. To have disadvantages, the tradition is to assume commuting costs for workers increase in a city, as city size grows and commuting distances increase, although the disadvantages can be expanded to include a variety of size-dependent disamenities (see below). All this is encapsulated in the monocentric city model, 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 the labour force is N, then the radius of the city is $\pi^{-1/2}N^{1/2}$. People living at distance b from the city centre spend t amount of working time to commute a unit distance (there and back), or face total commuting time costs of tb. 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 labour force of N, effective labour for a city is[1]

$$L = N - (2/3\pi^{-1/2}t)N^{3/2}. \tag{7}$$

This parameterization does not allow for congestion, so we experiment with and report results where t rises with N according to a constant elasticity form, so, in net, $L = N - (2/3\pi^{-1/2}\tilde{t})N^{z}$, where $z \geq 1{\cdot}5$.[2]

2.1.4. Net output per worker, city value-added, and city size. The model is solved in Appendix A. There, each final and intermediate good producer in the city chooses inputs to maximize profits. Then in the standard monopolistic competition framework, there is entry into local final goods and intermediate goods markets until profits are driven to zero at s_y final good producers and s_x intermediate good producers. These magnitudes are then related to total city employment through the local full employment condition. With these relationships we can then solve for the expressions for aggregate output and worker income.

The objective function we employ is net output per actual worker. This is the disposable income per worker in cities, after capital rentals are paid. If an individual city is to be of "optimal" (2nd best given our market institutions) size, it would want to maximize this magnitude, in a setting where there are many cities who compete for mobile workers in national labour markets (Duranton and Puga, 2004). Net output is city output less borrowing costs, or $(p\tilde{y} - rk_y)s_y$, where r is the rental cost of capital to the city. With various substitutions, from Appendix A, this equals $Q_2\mathrm{MP}^{1/(\sigma_y(1-\alpha))}r^{-\alpha/(1-\alpha)}\ A^{1/(1-\alpha)}\ L^{(\varepsilon+\gamma/\rho+\beta)/(1-\alpha)}$, where Q_2 is a parameter cluster.[3] Substituting in (7) for L, and dividing by N, we have

$$\text{net output per worker} = Q_2\mathrm{MP}^{1/(\sigma_y(1-\alpha))}r^{-\alpha/(1-\alpha)}A^{1/(1-\alpha)}(N - a_0N^{3/2})^{(\varepsilon+\gamma/\rho+\beta)/(1-\alpha)}N^{-1}. \tag{8}$$

In (8), for a given rental cost of capital, we can calculate the city size that maximizes net output per worker, or the size at the peak of the inverted U. Maximizing (8), net output per worker peaks at

[1] The formulation assumes land rents paid, which also rise with city size, are collected and redistributed in the city, as occurs in efficient free market solutions in these models—rent income paid out subsidizes the scale externalities optimally, a result called the Henry George Theorem (see Duranton and Puga, 2004). In China rent "redistribution" is more explicit, since land rents charged are nominal. In either case the resource cost is commuting time.

[2] Note any rise in unit commuting costs may be offset by increased density and smaller land consumption in bigger cities (where if lot size, h, is not normalized to 1, (7) is $L = N - (2/3\pi^{-1/2}t\ h^{1/2})N^{3/2}$). Note also that average commuting costs per person rise with city size, even as cities move from being monocentric to multi-centred (Fujita and Ogawa, 1982).

[3] $Q_2 \equiv Q_0^{1/(1-\alpha)}Q_1$, where $Q_0 = \sigma_y^{-1}(\sigma_y - 1)^{\alpha(1-1/\sigma_y)}c_y^{\alpha(1-1/\sigma_y)-1}\alpha^{\alpha}\rho^{\gamma}c_x^{-\gamma}\gamma^{\gamma/\rho}\ \beta^{\beta}(\gamma+\beta)^{-(\beta+\gamma/\rho)}$ $(f_x/(1-\rho))^{\gamma(1-1/\rho)}$, and $Q_1 = (1-\alpha)((\sigma_y - 1)c_y)^{(\sigma_y-1)/\sigma_y}$.

$$N^* = \left(\frac{\varepsilon + \gamma(1-\rho)/\rho}{a_0(\varepsilon + \gamma(1-\rho)/\rho + 1/2(\varepsilon + \beta + \gamma/\rho))} \right)^2 . \tag{9}$$

While N^* might be called "efficient" size, there are a variety of caveats concerning that label, which will be developed as the paper proceeds. We label N^*, "peak" size. Simple calculations show the following: (i) $\partial N^*/\partial\varepsilon > 0$. As city-scale externalities, ε, rise, peak size increases. (ii) $\partial N^*/\partial\rho < 0$. As substitutability, ρ, among intermediate inputs declines or the value of having more varieties increases, peak size increases. (iii) $\partial N^*/\partial\gamma > 0$ if $\beta(1-\rho) > \varepsilon\rho$. As the role of intermediate inputs, a sector with diversity economies, increases, or γ rises, peak sizes increase (with the parametric restriction ruling out a form of "super" scale economies by limiting how large the scale externality, ε, can be relative to labour's private return, β). As γ increases, if capital intensity, α, is constant (see later), β declines given $\alpha + \beta + \gamma = 1$; thus final output firms switch from internal labour usage (β declines) to local outsourcing (γ increases) to an intermediate sector where there are diversity economies.

We cannot estimate (8) directly, because in the data we do not observe r to calculate net output (and the implicit rental price may vary by cities in China's state-influenced capital markets); we only observe capital stock of the city and total city value-added, VA. Value-added of the city is $p\tilde{y}s_y$, which from Appendix A is given by $\text{VA} = Q_3 A\,\text{MP}^{1/\sigma_y} K^{\alpha} L^{\varepsilon+\gamma/\rho+\beta}$, where $K \equiv s_y k_y$, and Q_3 is a parameter cluster.[4] Thus,

$$\text{VA} = Q_3\, \text{MP}^{1/\sigma_y}\, AK^{\alpha}(N - a_0 N^{3/2})^{\varepsilon+\beta+\gamma/\rho}. \tag{10}$$

Given estimates of the parameters in (10), we can calculate the city size that maximizes net output per worker in (9), as well as assess how net output per worker in (8) varies with city size. Note also that, VA per worker from (10), $Q_3\text{MP}^{1/\sigma_y} A(K/N)^{\alpha}(1 - a_0 N^{1/2})^{\varepsilon+\beta+\gamma/\rho} N^{\varepsilon+\gamma(1-\rho)/\rho}$ given $\alpha + \beta + \gamma = 1$, is maximized at N^* in (9) for K/N held constant.

Different market allocation rules affect the exact form of (8) and (10). For example, in our derivation under monopolistic competition, the number of input varieties is not optimal as is well known. Optimality could be attained by paying per firm fixed costs from the local "public budget", something both China and the U.S. may approximate through local subsidy programmes. Under an optimal number of intermediate input varieties, (8) and (10) would look the same, except that the Q's would change. In that case, the institutional change only shifts the inverted U up or down, with no impact on its shape or the city size where the inverted U is maximized. Less "innocent" changes in institutions could, of course, affect the shape of the inverted U.

2.1.5. Manufacturing to service ratio. In estimation, one relationship will be of particular importance, since we use it to define how cities vary across the urban hierarchy when we generalize the model next in Section 2.2 to many types of cities. That relationship is the ratio of value-added in manufacturing to that in services, which we denote as MS. In Appendix A, we show $\text{MS} = (1-\gamma)/\gamma$, or

$$\gamma = 1/(1 + \text{MS}). \tag{11}$$

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, high-tech cities producing

[4] $Q_3 = Q_0\,\alpha^{-\alpha}(c_y(\sigma_y - 1))^{(1-\alpha)(\sigma_y-1)/\sigma_y}$.

scientific instruments or electronic goods, 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, Greenston and Healy (1972), and there are specific models detailing equilibria in such hierarchies.[5]

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

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

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 $\sigma_g = \sigma_y$, and only the consumption weights, μ_g, differ. The form to market potential in (6) now becomes more complicated, as discussed in Section 3.2.1.

The second element is to assume that aspects of production technology differ by product, so that there will be urban specialization by product (see below) and an urban hierarchy. In our data, we do not observe product specialization *per se*, but we do know the ratio of manufacturing to service value-added. In modern systems of cities, as we move up the urban hierarchy, the manufacturing to overall service ratio, MS, declines. In China the simple correlation coefficient between MS and city employment is about (−0·20), based on the overall service sector that is dominated by retailing and personal services, which tend to be a fixed proportion of GDP across all cities. For the U.S., 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 0·67 at the top size category.[6]

The manufacturing to service ratio identifies one parameter of the model in equation (11), where $\gamma_g = 1/(1 + \text{MS}_g)$. We implement an urban hierarchy in the model by setting $\gamma_g = 1/(1 + \text{MS}_g)$, so that MS_g tells us each city's value of γ_g. This relationship holds regardless of how other parameters vary across the urban hierarchy, and thus will also be the basis for describing the hierarchy in more flexible functional form approaches to estimating equation (10). The variation in γ_g is sufficient to give urban specialization. Ignoring inter-city transport costs of trade, specialization follows because, across the urban hierarchy, as γ_g rises and MS_g falls, from (9), peak city size increases. Having two different product types in the same city would result in a size that would be inefficient for at least one of the products. But inter-city transport costs are also a powerful force for integrating production of different products in the same city (Fujita *et al.*, 1999, ch. 11). To accentuate the forces for specialization, as is consistent with the empirical literature (see Rosenthal and Strange, 2004), it is common to assume that Marshallian-scale externalities, ε, are internal to the product, so that, for example, in textile cities, textile producers only learn from other textile producers (and their intermediate input suppliers) so equation (2)

[5] In the urban hierarchy literature, in a market economy with perfect migration, free capital markets, and developers and/or local governments involved in formation of new cities, any city type would operate 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 labour (as viewed by an individual city). As we move up the urban hierarchy, bigger cities have their peak points of 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 amenities vary within city types, then those with higher A^i's within a type operate to the right of their peak points in stable equilibria.

[6] As an illustration of this hierarchy, there are data on the spatial "product cycle", as reported in Fujita and Ishii (1999), on manufacturing electronics for the plants of big Japanese firms. Standardized production of generic television 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 occur in the largest metro area. Quite apart from the magnitude of information/knowledge externality issues, more experimental production requires more business services—outsourcing to programmers, designers, venture capitalists, advertising campaigns, etc.

becomes AL_g^{ε}. Then specializing by product type also maximizes external-scale benefits relative to urban commuting diseconomies.

Across the urban hierarchy if γ_g changes, given $\alpha+\beta+\gamma=1$, then either or both of β and α must change. Extensive experimentation in estimation led us to conclude α is invariant across the urban hierarchy in China. Thus, as γ_g rises and local outsourcing increases, manufacturers' use of labour, or β_g declines. In (10), the exponent of N, $\varepsilon+\beta+\gamma/\rho$, becomes $1+\varepsilon-\alpha+\gamma(1-\rho)/\rho = 1+\varepsilon-\alpha+(1-\rho)/\rho(1+\text{MS})^{-1}$. From that we get the basic equation that underlies all our estimation, structural or not. With substitutions, in logs equation (10) becomes

$$\begin{aligned}\ln \text{VA} = {} & \ln Q_3 + 1/\sigma_y \ln \text{MP} + \ln A + \alpha \ln K + (1-\alpha+\varepsilon) \\ & \times \ln(N - a_0 N^{3/2}) + (1-\rho)/\rho((1+\text{MS})^{-1} \ln(N - a_0 N^{3/2})). \end{aligned} \tag{10a}$$

We estimate two specifications of (10a). First, is a structural version, using the variation in MP, K, N, and MS in the non-linear specific functional form model in (10a) to identify $\sigma_y, \alpha, \varepsilon, \rho$, and a_0, the key parameters in assessing the inverted U. Structural estimation faces two issues. Empirical results in the literature suggest ε also varies across the urban hierarchy. For example, 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 U.S., finding a positive relationship. Thus, one might presume that the cluster $\varepsilon+\beta+\gamma/\rho$ in (10) as a whole rises across the urban hierarchy as both ε and γ/ρ rise. In addition, the exact form to urban commuting diseconomies may differ from what we have imposed, as noted earlier.

The second issue with direct estimation of the non-linear equation in (10a) is that, in principle, parts of the $\ln Q_3$ "constant" should vary across the urban hierarchy as MS varies (see footnotes 3 and 4). However, Q_3 identifies items such as how f_x varies with c_x, which are really beyond the scope of our aggregate data.[7] As a practical matter, we normalize $\ln Q_3$ to be a constant in estimation of (10a). Given these two issues, in assessing the exact shape to the inverted U and whether Chinese cities are undersized, we rely more on a version of (10a) where the terms giving the shape of the inverted U are collectively approximated by Taylor series expansions in MS and N, or transformations thereof, as discussed below.

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 (10a). Finally, we turn to estimation issues and results.

3.1. *Policy and cities*

3.1.1. Migration and urban policy. All migration in China is curtailed by the hukou system detailed in Chan (1994, 2000). Under the system, you are a "citizen" of the locality of which, traditionally, your mother is a citizen. Citizenship confers specific local benefits—access

[7] There are three components to $\ln Q_3$. First is a term $(1-\alpha-\gamma)\ln(1-\alpha-\gamma)$, where for 10–15% of observations with high values of $\gamma = 1/(1+\text{MS})$, $(1-\alpha-\gamma) < 0$ for typical estimates of α and $\ln(1-\alpha-\gamma)$ cannot be defined properly (except to impose a lower bound on $(1-\alpha-\gamma)$). Second, there is a parameter cluster, $\ln\rho + (1-\rho)\rho^{-1}(\ln(1-\rho) - \ln(1-\alpha)) - \ln c_x - (1-\rho)\rho^{-1} \ln f_x$, that multiplies γ where that cluster identifies how f_x varies with c_x in equation (2), given parameters α and ρ identified in other parts of the equation. (Once we have a variable $\gamma = (1+\text{MS})^{-1}$ with an "unconstrained" coefficient defining how f_x varies with c_x, we cannot anchor the $(1-\rho)/\rho$ coefficient in the last term of (10a), quite apart from the issue of how to deal with undefined $\ln(1-\alpha-\gamma)$ terms.) A third component, $\rho^{-1}(\gamma \ln\gamma)$, is no problem and utilizing it (constraining estimates of ρ to equal those in the last term of (10a)) leaves estimates reported below unchanged.

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 authorized about 18 million such changes a year from the early 1980's through 1997, and these involve a high proportion of urban–urban and rural–rural moves, rather than rural–urban moves underlying urbanization. For temporary migration to cities, 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 (legal or not) outside their permanent place of residence was under 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. While recent data and newspaper articles suggest a significant increase in migration in the last 5 years focused on a few cities, migration seemed in 1997 to be limited and mostly return, or round-trip migration. Rural–urban real income 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 (1981–1985). That plan, which continued in part to guide urbanization through the 1990's, intended to sharply constrain growth of large cities, while permitting limited migration through transfer of hukou from rural areas to towns and smaller cities. Evidence suggests that this planning combined with China's long-term aversion to large cities has distorted the size distribution of Chinese cities compared to other countries. Based on Henderson and Wang (2005), in 2000 China had only nine 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 0·072, compared to a worldwide ratio of 0·27. More generally, ranking cities by size from smallest to largest and calculating the cumulative share of urbanized population within a country, China's spatial Gini of 0·43 is substantially less than that for the world (0·56) and is smaller than all other individual large countries. Finally, we note that planners in the 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 foreign direct investment (FDI), with technology then "trickling-down" the hierarchy. We will want to account for this in estimation.

3.1.2. Market reforms. China from 1978 has undergone successive market reforms, as nicely summarized in Perkins (1994) for the period up to about 1990. These reforms 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 by the state, which occur primarily in 1993–1994. These reforms removed most of the remaining props under state-owned industry, exposing them to market competition. Most heavily hit were interior and northern heavy industry cities. These reforms moved most planning functions to a market basis and represent a break point in our urban data in terms of how outputs are evaluated. As part of the reforms moving into 1994 and extending into 1995, 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 detailed below, we will utilize this 1993–1994 split, viewing economic stock data in larger cities in 1990 as heavily determined by planning during the 1980's and flow economic data from 1997 as driven by market forces.

3.1.3. The urban system. We have data for 1990 and 1997 on about 225 prefecture-level cities (including four "provincial-level" cities). These are the larger formal cities in China, for

TABLE 1

Prefecture-level cities

	1990	1997	Growth (%)
Average population of the city proper (in thousands)	922	1087	18
Non-agricultural employment (in thousands)	415	527	27
Value-added per worker in non-agricultural sector (1990 yuan)	6389	10,588	66
Manufacturing to service (VA) ratio	2·17	1·44	−51

VA, value-added.

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 in the U.S. 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 205 cities used in the estimating sample (see Appendix B). From 1990 to 1997, their populations grew on average by 2% a year, but their non-agricultural labour 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 shorter-term immigrants and most 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 better capture urban expansion, and they are our size measure.

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 the freeing of most private business service activity in 1993–1994. In the data, over the 1990–1993 period the ratio declines modestly 1·5% a year; between 1993 and 1994 it declines by 24% (in part due to some redefinition of manufacturing as service activity in the reforms described above), and from 1994 to 1997, it declines by 4–5% a year. As restrictions on private service sector are removed, it takes off. By the late 1990's, while service growth continued, this decline dropped to about 2% a year. In estimation, we use 1997 data, in order to allow market forces the greatest opportunity to be fully operational, especially in the service sector. We do not use data after 1997 because the size measure, the total non-agricultural labour force, is no longer reported.

3.2. *Estimation*

3.2.1. Data and variables. A complete description of data sources and variables is given in Appendix B. Here, we note the highlights. Data in China are collected from the bottom up, by city statistical bureaus, following from the era when detailed economic planning governed allocations of factors and goods and involved a "twice up–twice down" process between the local and provincial planners. For prefecture-level cities, which each have their own statistical bureaus, the GDP data, at least up to 1997, are viewed as being extremely high quality. They are not subject to the same exaggerations experienced in recent years in the town and village enterprise (TVE) sector and are less likely to be manipulated, compared to manipulation at the level of the provinces and centre, creating the adding up problems that can exist in comparing national and local data. Price reforms in 1993–1994 as we noted above led to GDP evaluations based on market prices and eliminated any double counting that existed prior to reforms.

In terms of specific variables, 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. Labour force is the non-agricultural labour force. 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 collectives 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 (Au and Henderson, 2005). These controls are insignificant in our results, and have no effect 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 $\ln(\text{MP})$. For A we are looking for items that would affect the city-specific level of technology and labour force quality. We use the ratio in 1990 of people over age 6 with high school ("senior middle school") as a potential control for the 1997 labour force quality; we simply do not know 1997, age-relevant education attainment information for cities. For city-specific technology, we know accumulated (since 1990) real FDI by city (in U.S. dollars). We use the ratio of accumulated FDI divided by labour 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 did not 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. We also experiment with whether FDI levels in other nearby cities affect productivity following Bottazi and Peri (2003), or whether cities with better educated workers or higher up the urban hierarchy benefit more from greater FDI.

Finally, we need to construct a measure of market potential, $\ln(\text{MP})$. There is a trade literature, which attempts to estimate the elements making up this variable, based on trade-flow information, industry by industry across many country pairs (*e.g.* Overman *et al.*, 2003; Hummels, 2004). We do not have the trade-flow data to do this. Similarly Hanson (2005) for the U.S. infers key elements of market potential; but to do so he needs to assume perfect labour mobility across U.S. counties and utilize both wage and income information. One key point of our estimation is that labour is highly *im*mobile in China, invalidating the use of such an approach; besides, we do not have the required wage data. Instead, what we need to do is construct a measure of market potential for each of our cities in China *a priori*, utilizing results from this trade literature. In calculating such an index for market potential as in equation (6), there are five issues.

First is how to measure expenditures E_υ in localities. For that we use total GDP of the whole prefecture (not just the urbanized area). These prefectures cover most of China, but we supplement them by adding in county cities outside the control of these prefectures ("under the province") as units to ship to, since their GDP is not counted in prefecture GDP. The second issue concerns how to distance discount, to represent how transport costs rise with distance. The literature (*e.g.* Hummels, 2004) assumes a function for the unit transport cost factor, $\tau_{j\upsilon} = Ad_{j\upsilon}^{\delta}$ where $d_{j\upsilon}^{\delta}$ is the distance from the centre of locality j to that of υ, where then this function is raised to the power $1-\sigma_y$ in (6). Hummels estimates the elasticity δ for rail traffic for the U.S. as being 0·57. For China with its slower and universally utilized rail system, Poncet (2005, table 1, column 10) estimates a value of 0·82 for δ. In the trade literature using aggregate data (as opposed to detailed sector data), typical values of σ_y are about 2. For example, while Poncet's numbers for σ_y bounce around, for the δ of 0·82, her corresponding σ_y is 1·6. *A priori* we felt this was a little low and set $\sigma_y = 2$. As it turns out, results are not sensitive to the exact choice of σ_y in the neighbourhood of 2, and we will obtain our own estimate of σ_y to compare with the assumed value in discounting.

If in equation (6), we distance discount by $(Ad_{j\upsilon}^{0\cdot 82})$, what is the value of A in the calculation? That raises another issue: how to calculate d_{jj} the distance for the own city. For that, the standard procedure (*e.g.* Davis and Weinstein, 2001) is to use the average distance travelled by consumers in a city to shopping in the city centre (again assuming fixed lot sizes and a circular city), which is two-thirds the radius of the city. The radius is for the whole prefecture and all distance units are in hundreds of miles. For A, we choose the value such that $(Ad_{jj}^{0\cdot 82}) = 1$ for the smallest land area city in the sample, noting that d_{jj} is 2/3 the radius of that city, or $2/3\pi^{-0\cdot 5}\text{area}^{0\cdot 5}$.

Fourth and most troubling are the I_{υ} in equation (6). First, we note with multiple types of products as in (4a), market potential is product specific, where the function in (6) is multiplied by a product share coefficient[8], and critically, the I_{υ} are product specific, referring to the gross prices within the same product group for all cities that produce that product imported by city υ. Not only do we not have price data by city, we cannot assign what cities produce what products. In calculating the measure of market potential, we have little choice but to normalize all I_{υ} to be 1, so we are using what is called nominal market potential, instead of real market potential (Head and Mayer, 2004). To try to capture possible biases from doing this, we will experiment with interacting the calculated market potential measure with various variables, such as the own city's MS ratio, and latitude and longitude where geographic patterns of production vary north to south and east to west. The interaction with the MS ratio represents the city's own product type and could also help correct for the fact that we have considered only consumer and not producer markets for inter-city traded good products. In principle, one could specify an estimating model with separate intermediate input demand for products; but again we do not have the data to estimate such a specification.

The final issue is how to deal with international income, E_R, where in the transport cost factor $d_{j,\text{coast}}$ is distance from city j to the China coast. To incorporate E_R, we decompose ln(MP) in (6) into

$$\begin{aligned}\ln(\text{MP}_j) &= \ln\left(\sum_{\upsilon\in\text{China}} \frac{E_{\upsilon}}{(Ad_{j\upsilon}^{0\cdot 82})} + \frac{E_R}{(Ad_{j,\text{coast}}^{0\cdot 82})}\right)\\ &\approx \ln(\text{MP}_{j,\text{domestic}}) + E_R \frac{1}{\text{MP}_{j,\text{domestic}}\ (Ad_{j,\text{coast}}^{0\cdot 82})}\end{aligned} \tag{12}$$

where $\text{MP}_{j,\text{domestic}} \equiv \sum_{\upsilon\in\text{China}} \frac{E_{\upsilon}}{(Ad_{j\upsilon}^{0\cdot 82})}$. The first-order Taylor series expansion approximation in (12) assumes that the domestic component of market potential is very large relative to the international component for most cities in China, as our results will suggest. Note that in (10a), ignoring issues with our calculations of market potential, $\ln(\text{MP}_{j,\text{domestic}})$ in (12) has a coefficient of $1/\sigma_y$, while the second term has a coefficient of E_R/σ_y, which potentially allows us to identify E_R, to compare foreign with domestic market potential for cities.

3.2.2. Econometric issues. A key issue is the error structure for equation (10a). In general in a market context, there are unobserved variables that affect productivity and hence input choices. For example, current shocks to city productivity, such as a recent import or adaptation of a new technology may affect city investment and wages, inducing in-migration. Time-persistent shocks to do with unmeasured location–geographic features or local political and institutional environments again may affect both productivity and factor allocations. Finally, variables may be measured with error, resulting in attenuation bias.

Our strategy to deal with these problems affecting identification is to instrument for *all* 1997 time-varying covariates with historical characteristics of the city and estimate equation (10a) by

[8] In estimation given the log form to (10a) this is a constant by-product, which is subsumed in the error term. There is no reason to expect these product demand magnitudes to be correlated with technology ones.

non-linear 2SLS (two-stage least squares) and its flexible functional form version by 2SLS. Because current magnitudes present accumulation processes (see below), our instruments are strong, with first-stage F's and R^2's averaging over 65 and 0·75, respectively. The issue is their validity, or exogeneity to current shocks affecting current productivity. In this section we articulate an economic rational for choices of specific instruments, and then we turn to the practical aspect—tests for such exogeneity. Details of results on specification tests and first-stage regressions are posted on the journal website, along with the data used in this paper. The economic rationale for choice of instruments has two parts, each relating to a specific set of historical variables.

Planning variables. The 1990 capital to labour ratio, percentage of population over 6 with high school, spatial area of the central business district (and that interacted with the manufacturing to service ratio), agriculture to other sector ratio, FDI to labour force, sales of independent accounting units to all enterprises (see data section) and whether a city had FDI are used as instruments, as variables influenced largely by planning and politics, and exogenous to unobservables affecting productivity in 1997. The key assumption is that provincial planners in the 1980's in making allocations to cities that give us these 1990 variables ignored unmeasured (to us) aspects of the local environment, which affect productivity in both 1990 and 1997. The argument is based on certain facts and one assumption. First, in the late 1980's, for prefecture-level cities, unlike the rural sector and in smaller cities, these variables were still largely determined by planners and government officials. FDI, for example, was explicitly controlled and vetted, with designated FDI sites. Second, planners and officials' objectives were not to make allocation decisions to maximize the market value of output per worker *per se*, but rather to satisfy certain planning and political objectives, although planning objectives would encompass aspects of productivity. But to the extent productivity entered the planning calculus, the assumption is that final decisions by provincial planners were based on the same observables we have access to, and not on the unobservables, at least ones persisting in impact to 1997. Finally, we note that to the extent managers of state-owned urban firms in the 1980's had autonomy, managers were heavily restricted by local politics. They operated with a limited idea of how to respond to market forces and had a very limited incentive to do so—stated owned firms operated with no hard budget constraint.

As we move into the 1990's, reforms in the state-owned urban sector are less cosmetic and more cutting. In particular in 1993–1994, the state-owned sector is moved in a dramatic fashion to a market basis and the service sector, particularly business services, as noted earlier is freed up with vast expansion of private services. We perceive this as a regime switch, where many of our cities between 1993 and 1994 stretching into 1995 have enormous changes (up or down) in Y/N, K/N, and MS for roughly the same urban scale, indicating dramatic shifts in the way quantities were evaluated. By 1997, we perceive economic magnitudes driven primarily by market forces.

In summary, the first rational for instrumenting is that there are a set of variables from 1990 reflecting planning decisions in the 1980's, which are strong instruments but are unaffected by unobservables affecting productivity in 1997. However, in testing (later), the absolute size of the city labour force or output in 1990 are not exogenous. Planning *ratios*, or planning coefficients like capital per worker are, but not *absolute scale*.[9] Even in 1990, to the extent possible, migrants may have responded to unobservables (to us) affecting productivity and potentially earnings.

Amenity variables. For labour force currently and historically we have a different instrumenting rational, which is based on migration decisions and parallels the classic case of using demand variables to instrument for price in estimating supply curves. The model for this is given

[9] Of course, in some cases even in a market context, unobservables could affect the scale of variables in the same proportion and hence not affect their ratios.

in a companion paper, Au and Henderson (2005), in some detail, and here we briefly summarize it. In China, as discussed earlier, migration to cities is very costly and most migration up to 1997 is local, from, in particular, the rural parts of a municipality into its city proper.

To model this, we assume a "demand side" where each city offers a utility level to a resident, which is a function of its real wage and local quality of life, Q, where Q are consumer amenities potentially distinct from producer amenities, A. Real wages are related to the city's allocation of labour, and capital, as well as technology, so the city utility that can be offered to migrants is some function $U = U(K, N, A, Q)$. The "supply side" comes from the local rural sector, where utility is similarly determined by rural capital, labour and consumer and producer amenities, or $R = (K_R, \bar{N} - N, A_R, Q_R)$, where $\bar{N}$ is the total population of the whole local region of the city. If there were no migration restrictions N would adjust to equalize R and U. But in China, we presume $U > R$ 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, $\dot{N} \equiv (dN/dt)/N$, rises. At any instant the gap between urban and rural utility equals the cost of migration $m(\cdot)$, or $U_t - R_t = m(\dot{N}), m, m' > 0$. Such an equation is the specification of migration frictions in the U.S. (in part due to rising costs of housing with in-migration in the short run), used in Mueser and Graves (1995) and Rappaport (2000). It provides a link through migration accumulations between current city employment and a historical rural base population, and it provides a link between city amenities and migration accumulations.

To instrument for the current labour force, for 1990, we assume there is a (large) rural–urban utility gap based on historical allocations within the municipality in question. We take the 1990 population of the rural area as exogenous and the base for much of the migration into the nearby city determining 1997 labour force. And we have measures of consumer urban amenities in 1990, which we presume are related to 1997 amenities. These are library books, doctors, telephones, and roads, all *per capita*. These amenities along with the measure of surrounding rural population, we call amenity instruments. An issue might be whether 1990 (planned) consumer amenities might also reflect unmeasured production amenities in 1997, but we test for their exogeneity.

3.2.3. Specification tests. The full instrument list of planning and amenity variables yield very good specification test results for all models. We initially performed informal tests such as (1) pooling 1995–1997 data to estimate city fixed effects and then regressing these fixed effects on instruments to determine that instruments are uncorrelated with the estimated fixed effects[10] and (2) including instruments along with our covariates in ordinary estimation to ensure instruments don't affect covariate coefficients. These informal tests all yield good results for the instruments we use. For formal tests, we rely on χ^2 tests on over-identifying restrictions, based on the R^2 from regressing residuals from IV estimation on instruments. For these, in all the models, no individual instruments in the residual regressions are close to having significant coefficients and the χ^2-test statistics reported in the tables below are well within the acceptable range. But if we add to our instrument list the excluded absolute labour force (or, total value-added) in 1990, specification test results fail.

3.3. *Results for the structural model*

In this section, we report key results for the structural model and look at net urban agglomeration economies for that model. Then we turn to more flexible functional form models. For the

[10] However, we reject a fixed effects approach *per se*. First we do not think fixed effects are the correct error structure (as opposed to an AR process). Second changes in covariates from 1995 to 1997 are noisy, in part because of ongoing price and economic reforms that change valuations and shock sectors stretching through 1995 and even 1996. Finally, attenuation bias from measurement error is accentuated in fixed effects and our instruments are weak for changes in magnitudes (as opposed to levels).

TABLE 2

Results for urban productivity (*S.E. in parentheses*)

	IV estimation structural model	Ordinary non-linear least squares structural model
a for capital	0·428** (0·0846)	0·417** (0·0442)
$(1-\alpha+\varepsilon)$	0·605** (0·182)	0·576** (0·874)
$(1-\rho)/\rho$	0·425** (0·187)	0·143* (0·0779)
$-a_0(=2/3\pi^{-1/2}t)$	−0·0347** (0·00494)	−0·00833 (0·0228)
% High-school education	0·000473 (0·00432)	0·00432 (0·00313)
FDI per worker	0·0793** (0·0272)	0·0727** (0·0166)
$1/\sigma_y$	0·650** (0·0987)	0·536** (0·0790)
E_R/σ_y	1·46 (2·91)	4·45** (2·01)
Constant	0·182 (1·13)	1·38* (0·741)
N	205	205
R^2	0·914	0·923
χ^2-test statistics from specification test (critical value)	14·8 (16·9)	

FDI, foreign direct investment.
*Significant at 10% level.
**Significant at 5% level.

structural model here, we focus just on the scale effect results, as well as capital intensity, delaying discussion of market potential and technology variables to Section 3.4.

Basic results for the non-linear 2SLS estimates of coefficients are given in Table 2, column 1. Regular non-linear least squares results for this model are given in column 2 for comparison. The main effect of IV estimation is on scale variable coefficients, reflecting the problem noted earlier of endogeneity of migration responses. In order to interpret all results we start by examining the results on capital intensity. In Table 2, column 1 the coefficient, α, is 0·43. This high coefficient does drop somewhat for the flexible functional form approach to (10a) below, but a high coefficient is 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). In various specifications, interacting the capital variable with employment scale, education, or the manufacturing to service ratio results in small insignificant effects for the interacted variable, leading us to conclude capital intensity does not vary across cities.

3.3.1. Scale economies. There are two parameters essential to identifying scale economies in equation (10a). The first concerns diversity scale effects. From the discussion of equation (1) where y collapses to $A(\cdot)k_y^{\alpha}\,\ell_y^{\beta}(xs_x)^{\gamma}s_x^{\gamma(1-\rho)/\rho}$, and from the expression for s_x in Appendix A, a 1% increase in effective labour leads to a $\gamma(1-\rho)/\rho$ per cent increase in city output. In our sample, MS takes a typical (average and just above the median) value of 1·4, for which $\gamma = 0{\cdot}42$. Given $(1-\rho)/\rho = 0{\cdot}425$ in the table, for the typical city this implies a city-scale elasticity due

TABLE 3

Urban agglomeration: city employment at the peak to net output per worker

MS	0·6	1·0	1·4	1·7	2·0	2·5	3·0	4·0
Peak point in thousands	1441	1174	1019	926	849	744	663	544
Lower* 95% confidence interval	977	749	552	411	283	99		
Upper 95% confidence interval	1905	1598	1486	1441	1414	1390	1376	1360

*A blank indicates a negative lower bound.

to diversity effects of 0·18. This is very high, indicating the forces behind the size of large metro areas that have high concentrations of services—returns to diversity in service activity. Note $(1-\rho)/\rho = 0{\cdot}425$ implies $\rho = 0{\cdot}702$ so the elasticity of substitution in production among intermediate inputs is 3·4. This seems a reasonable number for products defined at this level of aggregation.

The second-scale economy in equation (10a) is the degree of Marshallian-scale externalities ε. Given $\alpha = 0{\cdot}428$ and $1-\alpha+\varepsilon = 0{\cdot}605$, that implies an elasticity $\varepsilon = 0{\cdot}034$, which seems low; but is plausible for aggregate manufacturing. An ε of 0·033 says that a 10% increase in the local labour force increases productivity by 0·33%, a typical mid-range estimate of ε across disaggregated manufacturing industries (see Rosenthal and Strange, 2004, for a review of studies). However, this elasticity is not significant, having a S.E. of 0·109. We experimented with the functional form to scale externalities having ε decline with scale (so the exponent becomes $\tilde{\varepsilon}/N$), but that also produced insignificant results.

In summary, the results for service oriented prefecture-level cities at the top end of the urban hierarchy in China suggest scale diversity effects are the dominant source of agglomeration benefits.

3.3.2. Urban diseconomies. In equation (10a), the coefficient on $N^{1{\cdot}5}$, which is a_0 in Table 2 gives total commuting costs of $0{\cdot}035\,N^{3/2}$. For a typical city with a labour force of 500,000 and a population of 1·0 million (where population is typically twice the labour force), that implies 12·4 of the labour force of 50 (for N in units of 10,000), is used up in commuting activity—about 25%. This seems high but not unreasonable in a developing country, defining commuting broadly to include all commuting, such as the extra time devoted to local work and school-related trips and shopping time as city sizes increase. The United Nations Centre for Human Settlements (UNCHS) data for 1996 on world cities suggest that about 15% of work time is spent just on the commuting to work trip. We did experiment with other exponents, z in $a_0 N^z$, to allow "congestion" as discussed earlier. As we raise the exponent z, trying to accelerate how commuting costs rise with city size, surprisingly, the a_0-coefficient multiplying that covariate falls so much that, the proportion of time spent commuting in cities *declines* as the exponent rises. For example, for an exponent of 1·7, the a_0-coefficient falls from 0·0347 to 0·00957. Then for a city of 1 million population and 500,000 workers, the fraction of time spent commuting falls from 25% to 15%. Correspondingly, with this reduction in commuting costs under higher values of z, even more Chinese cities would be assessed as being undersized in Section 4.

3.3.3. Net urban agglomeration economies. Table 3 illustrates the peak sizes of cities for this specification, using the coefficients in Table 2 to calculate peak sizes in (9). But we should be clear about one simplification. As city sizes increase, in calculating peak sizes, we are only considering the internal-scale economies and diseconomy effects used to calculate N^* in equation (9). There is another scale effect, which we hold constant, because it really isn't feasible to calculate the required full general equilibrium feedback effects city by city in the specific Chinese geography. This effect has two components. First, as the size and GDP of a city expand, that expands its own market potential—that is, the city, as well as exporting, buys from itself raising its own demand. Second there is the virtuous economic geography feedback where as a city expands that increases its demand for other nearby city's products that increases their GDP, which feeds back into the first city's market potential. By holding constant market potential, we are potentially understating scale effects and peak sizes; but that only reinforces our results that in the end Chinese cities in 1997 are too small.

Table 3 shows how the peak points vary, as the manufacturing-to-service ratio varies. The results we discuss here are similar to those obtained next for flexible functional form models. 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 ($\text{MS} = 0{\cdot}6$) have peaks at an employment of 1·4 million or population of 2·8 million. This may seem small, given the sizes of modern metropolitan areas in the world, but few Chinese cities are in this range. While the stated objective of larger Chinese cities is to have $\text{MS} < 1$, less than 24% of prefecture-level cities met that objective in 1997; and only six cities have MS values less than 0·6, 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 just over and a median just under 1·4 and 18% of cities have values in excess of 2·0 where the peak size is at employment of 0·74 million. A few Chinese cities remain extremely manufacturing intensive with MS values ranging up to 4. At high MS values, the employment values for the peak point tail off.

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 the wide confidence intervals for this specification and the ones to follow. Actual city sizes lie both to the left and right of the point estimates of where the peaks lie, but with only 10% of the 205 cities having actual sizes to the right of their peak. This is in marked contrast to what is expected in a free migration economy—virtually all cities being to the right.

3.4. *Results for flexible functional forms*

In this section we examine the shape to the inverted U of net output and VA per worker against city employment, giving a more flexible form to (10a) and focusing on net agglomeration economies, without trying to separate out the components. In addition to net agglomeration economies, we examine the effects of technology and market potential on output per worker. To clarify the distinction for flexible functional form models between net output vs. VA per worker, following equation (10a) specified in VA per worker form, we always estimate VA per worker relationships. To get the shape to the inverted U for net output per worker, from the analysis of equations (8) and (10), as employment changes, $\partial \ln(\text{VA}/N)/\partial N = (1-\alpha)\partial \ln(\text{net output per worker})/\partial N$, where on the R.H.S. the capital rental rate is held fixed and on the L.H.S. the capital to labour ratio is held fixed. Given that, the shapes to VA per worker vs. net output per worker are the same up to a factor of proportionality; and both peak at the same employment size. In discussing the shape to the inverted U of net output per worker, we will convert estimates for VA per worker using the $(1-\alpha)$ factor of proportionality.

3.4.1. Net agglomeration economies. Before doing econometric estimation, we first plotted a graph of the raw data for VA/N against N, which, while hinting at overall modest inverted U, 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 and in footnote 5, 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 (see later). 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 labour markets. China does not have free migration; but there is no particular reason to expect a specific shape to the plot. 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. Initially, to see whether such a relationship might exist we combined data for 1996–1997 to increase sample size and broke the sample into septiles based on MS values. We then did OLS regressions of value-added per worker, against basic covariates with a quadratic in N and calculated N^* for each MS interval. For the lowest MS septile, N^* is at 2·3 million workers. At the second, it jumps to 4·3 million, but then after it declines monotonically taking values, respectively, of 2·4 million, 1·4 million, 1·3 million, 0·60 million, and 0·28 million. At the upper end these are larger city sizes than we find empirically, but OLS results generally show larger peak sizes than IV estimation. Having gotten suggestive results we then turned to detailed econometric work.

Estimation is based on (10a), where we start with

$$\ln(\text{VA}/N) = 1/\sigma_y \ln \text{MP} + \ln A + \alpha \ln(K/N) + [\ln Q_3 + (\beta + \gamma/\rho + \varepsilon) \times \ln(N - a_0 N^{3/2}) - (1-\alpha)\ln N].$$

For the term in square brackets, while γ and β vary directly with MS, we expect ε to vary across the urban hierarchy and commuting costs to take a more complex relationship than engendered in (7). To capture this, we approximate the expression in square brackets by a second-order Taylor series expansion in MS and N to get

$$\ln(\text{VA}/N) = 1/\sigma_y \ln \text{MP} + \ln A + \alpha \ln(K/N) + [a_1 N - a_2 N^2 - a_3 N \times \text{MS} + a_4 \text{MS} + a_5 \text{MS}^2]. \tag{10b}$$

While we report results on this expansion in MS and N, for reasons discussed below, we prefer a generalized Leontief form, where the second-order expansion is in square roots. We tried third-order expansions, but given the limited sample size and multicollinearity inherent in higher-order expansions, third-order expansions have insignificant coefficients for all expansion terms.

In (10b) the presumption is that $a_1, a_2, a_3 > 0$, and $a_1 - a_3\text{MS} > 0$. Maximizing value-added per worker, holding constant the K/N ratio, gives a peak size of

$$N^* = \left(\frac{a_1 - a_3\text{MS}}{2a_2}\right). \tag{13a}$$

For the expansion in square roots, for the corresponding parameters, peak size is

$$N^* = \left(\frac{a_1 - a_3\text{MS}^{1/2}}{2a_2}\right)^2. \tag{13b}$$

Results on net agglomeration economies. Results for equation (10b) and its version with an expansion in square roots are given in Table 4. Column (1) is for the generalized Leontief and (2) for the regular Taylor series expansion, both estimated by 2SLS. We start with a discussion of capital intensity and net urban-scale externalities and then turn to technology and market potential variables. For capital intensity the coefficient, α, now is more within international norms taking a point estimate of 0·36 in both columns. For net-scale effects, the coefficients on the Taylor

TABLE 4

Flexible functional form specifications

	IV estimation	IV estimation
	Generalized Leontief	Regular Taylor series (terms in square brackets)
$\ln(K/N)$	0·362** (0·0916)	0·363** (0·0897)
$N^{0\cdot5}\,[N]$	0·366** (0·116)	0·0102** (0·00230)
$N\ [N^2]$	−0·00805** (0·00254)	−0·0000140** (0·00000394)
$N^{0\cdot5} \times \mathrm{MS}^{0\cdot5}\,[N \times \mathrm{MS}]$	−0·184** (0·0872)	−0·00474** (0·00199)
$\mathrm{MS}^{0\cdot5}\,[\mathrm{MS}]$	0·218 (1·93)	−0·128 (0·278)
$\mathrm{MS}\,[\mathrm{MS}^2]$	0·206 (0·615)	0·0508 (0·0521)
% High-school education	0·00142 (0·00491)	0·00209 (0·00452)
FDI per worker	0·0683** (0·0286)	0·0652** (0·0291)
$\ln(\mathrm{MP}_{j,\mathrm{domestic}})$: $\{1/\sigma_y\}$	0·680** (0·117)	0·746** (0·109)
$(\mathrm{MP}_{j,\mathrm{domestic}}(Ad^{0\cdot82}_{j,\mathrm{coast}}))^{-1} : \{E_R/\sigma_y\}$	3·94 (3·16)	3·94 (3·28)
Constant	0·00576 (1·35)	0·593 (1·01)
N	205	205
R^2	0·550	0·530
χ^2-test statistics from specification test (critical value)	10·8 (16·9)	10·3 (16·9)

FDI, foreign direct investment.
*Significant at 10% level.
**Significant at 5% level.

series expansions have no structural interpretation, but we do note that the two MS terms, which could be thought of as controlling for the Q_3 term in (10a), have insignificant coefficients in both expansions.

For results on net-scale economies we turn to Tables 5 and 6 and Figure 1. Table 5 gives the peak points where VA per worker (and also net output per worker) is maximized. For column (1) point estimates in Table 4, 18% of cities are to the right of their peak points and the rest to the left, while for column (2), 21% of cities are to the right. In Table 5, as in Table 3, peak points decline as the MS ratio rises. Our preference for the generalized Leontief in Tables 4 and 5 derives in part from the fact that for it we can calculate peak points for all but two data points in the sample, whereas a regular Taylor series expansion has no peak points for MS values in excess of 2·1, where about 15% of cities have such values. Compared to the structural model, the flexible functional form models show larger peak sizes for more service oriented cities but smaller sizes for intensive manufacturing cities. Table 5 also shows the 95% confidence intervals for peak sizes. While the results in Table 3 and for the two versions in Table 5 differ in calculations of peak

TABLE 5

Urban agglomeration: city employment at the peak to net output per worker

MS	0·6	1·0	1·4	1·7	2·0	2·5	3·0	3·5
Case 1: generalized Leontief								
Peak point in thousands	1919	1270	842	607	426	213	83	17
Lower* 95% confidence interval	1162	984	415	659				
Upper 95% confidence interval	2675	1557	1268	1148	1025	797	540	260
Case 2: regular Taylor series								
Peak point in thousands	2624	1946	1269	760	252			
Lower* 95% confidence interval	2073	1617	635					
Upper 95% confidence interval	3175	2276	1902	1730	1577			

*A blank indicates a negative lower bound.

points, most of the difference is in the tails. For the median MS value around which most cities lay, 1·4, the models give similar results. For Table 3, and case (1) and (2) of Table 5, respectively, the 95% confidence intervals for MS = 1·4 are 552–1486, 415–1286, and 635–1902. In all, as we will see in Section 4, the different models show similar numbers of significantly undersized cities, although the results we favour most for case (1) of Table 5 show the fewest significantly undersized cities.

Table 6 and Figure 1 illustrate variations in value-added per worker as cities move away from their peak sizes. Figure 1 shows net output per worker for MS = 1 and for an extreme value of MS = 2·7, both for net output per worker normalized to 18,000 yuan per year at the peaks to the inverted U. Table 6 shows the deviations in output per worker as size moves away from peak size for MS = 1. The calculations are based on the column (1) coefficients in Table 4, although the qualitative results on the asymmetry of effects of being oversized vs. undersized are the same for estimates based on Table 2 or column (2) of Table 5. For the column (1) coefficients in Table 4, Table 6 gives the per cent loses in net output per worker from operating at a size away from the point estimate of the peak, calculated as

$$\ln(\text{net output}/N)^* - \ln(\text{net output}/N) = \frac{1}{1-\hat{\alpha}}\{(\hat{a}_1 - \hat{a}_3\text{MS}^{0\cdot5})[(N^*)^{0\cdot5} - N^{0\cdot5}] - \hat{a}_2(N^* - N)\} \tag{14}$$

where an asterisk is the value at the peak. The ratio MS is held constant. As usual in absolute value terms, (14) is the same approximation for both losses of moving away from the peak and gains of moving to the peak.

Several things are apparent in Table 6 and Figure 1. First, there are enormous agglomeration economies. Moving from a city with a labour force of 100,000 to 1·27 million for MS = 1 raises real output per worker by 83%, 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 635,000 to 1·27 million only increases real output per worker by 14%. This notion is more explicitly explored in Table 6 for MS = 1, which shows the per cent of current output per worker to that at the peak, as a city moves away to the left and right from its peak size. Third, in Figure 1 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 smaller than being undersized. For MS = 1, decreasing vs. increasing city sizes by 50% reduces net output per worker by 14% vs. 8%. Or from a peak size of 1·27 million if one subtracts 1·22 million people so city size is 50,000, real output per worker falls by 83%; while, if one adds 1·22 million so size is 2·49 million, it only falls by

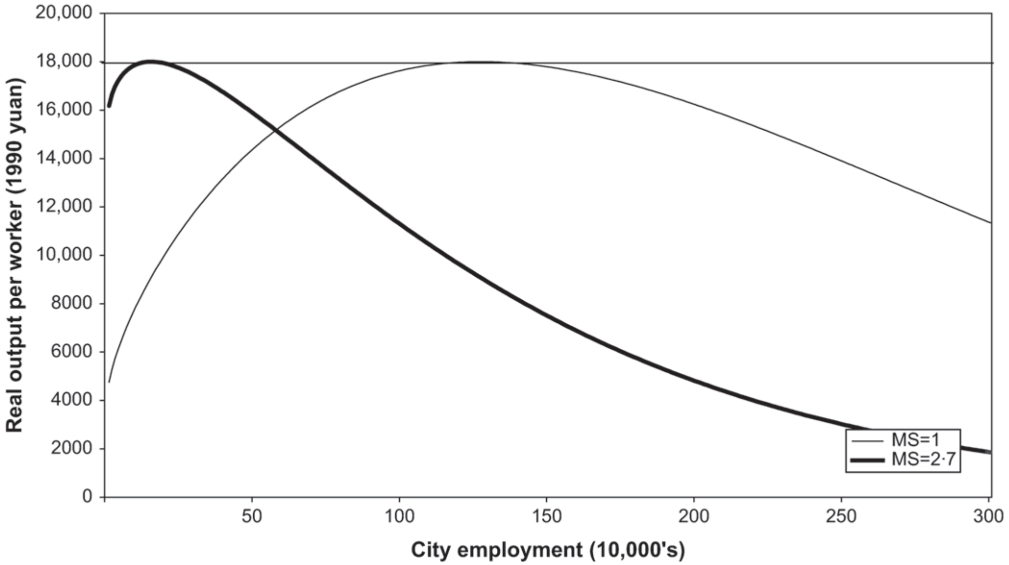

FIGURE 1

The inverted U for cities

26%. Real output per worker has a fairly flat portion near the peak, and real output per worker initially drops slowly past the peak. This has implications for free market analysis of city sizes, with differing amenities across cities. Among cities of the same type, those with better market potential or amenities have their inverted U's shifted up. With free migration in Figure 1 and, say, a horizontal supply curve of people at 18,000 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 18,000 at the same size (1·27 million). 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 18,000. Given that real output per worker declines fairly slowly past the peak, this could be a very large size.

Tables 5 and 6, as well as Table 3, also have 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. Finally as discussed above, what is "constrained optimal" in a world of perfect mobility and heterogeneous local urban amenities is for most cities to be to the right of their peak points, although 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 country like China, with an essentially uncountable number of viable urban sites, it is unclear how much natural amenity differentials across urban sites really matter. What is clear is that free migration would result in large increases in city sizes and productivity gains.

3.5. *Other results*

Finally we have the results on technology and demand variables in column (1) of Table 2 and columns (1) and (2) of Table 4. The results are all similar and we focus on those for column (1) of Table 4. We start with market potential. The coefficient on domestic MP can be interpreted as an estimate of $1/\sigma_y$, so $\sigma_y = 1{\cdot}5$, which is the elasticity of demand for a city's product, with the caveat that it is based on a measure of nominal not real market potential as discussed

TABLE 6

Agglomeration benefits (MS = 1)

	Employment in thousands										
	20	50	100	320	635	950	1270	1590	1900	2490	3000
Per cent gain in net output per worker of moving to peak size $N^* = 1270$	133	103	83	40	14	2·9	0	2·3	8·0	26	46
Current city size as a per cent of peak size	1·6	3·9	7·9	25	50	75	0	125	150	196	236

earlier. For aggregate data, this corresponds to results in the literature, recalling for example that Poncet's estimate of σ_y is 1·6 for her value of distance discounting that we use. Given the S.E. on the estimate, its 95% confidence interval easily encompasses the value of 2 for σ_y we used in constructing the market potential measure. Apart from this structural interpretation, the result tells us that a 1% increase in market potential for a city leads to a 0·16% increase in value-added for the city. Local regional demand is a critical component of measured city productivity.

Because of the issue of real vs. nominal market potentials we also tried interacting the market potential measure with MS and latitude and longitude, to represent both the possibility that there is demand for city manufactured products as intermediate inputs nearby and the fact that regional patterns of production vary from north to south and east to west as natural resources vary. None of these interacted variables are significant. They have small coefficients and the coefficient on lnMP itself is unchanged by inclusion of these terms.

Market potential considerations include an international demand component, represented in equation (12) by market potential with distance to the coast. That variable is positive but never significant. If we replace it with a dummy variable for being a coastal city and by distance to the coast for non-coastal cities, these variables are similarly insignificant. But as we note below the magnitude of the coefficient is not trivial; it is just that the S.E. is large. For idiosyncratic reasons, certain cities in China have become very export oriented while others have not. For example certain cities in China were developed as official "export zones" ("open cities" and the like). However, a dummy for these favoured cities is insignificant with no change in the international market potential term, which indicates that these favoured cities are not inherently more productive than other cities (controlling for their FDI and capital stock levels). As to the magnitude of the point estimate, first we note that average domestic market potential for cities is about 1452 units. For the international variable, the coefficient equals E_R/σ_y, but the variable is normalized (multiplied by 1000). Accounting for the normalization and the base value of A (15·3) in discounting in (12), gives an international market potential of 378 for a coastal city in 1997. For a typical city further from the coast, the average discount factor is 40. So the international market potential for the "typical" city is 145, compared with the domestic number of 1452.

For technology variables, there is education and accumulated FDI per worker in thousands of dollars. For the latter, a one-S.D. increase leads to a 11% increase in VA, a very large effect. Technology transfer through FDI seems to bring high productivity benefits, perhaps a justification for why the Chinese subsidize these transfers. Cities favoured by policy in the early 1990's as FDI targets gained a significant advantage. We also looked into the issue of FDI spillovers across cities, controlling for FDI in nearby cities, within 150 miles. In Europe, Bottazi and Peri (2003) find high spillovers within the own area, but also very small but significant spillovers from immediate neighbours (only). In China with its poorer communications, we found no evidence of any spillovers from near neighbours; this is consistent with the evidence on the quick spatial decay of information spillovers in the U.S. (Rosenthal and Strange, 2004).

Finally, there is education. Unfortunately, we do not have 1997 values and have to rely on 1990 values. Moreover, these measures are for the population age 6 or over completing high school, not just for adults. The coefficient is essentially 0. Interactions of this variable with scale variables or FDI produced insignificant, small effects. The zero coefficient is disappointing but we ascribe the result to having a poorly measured variable. We also note we are looking at prefecture-level cities where education levels are fairly uniformly high, given migration restrictions that funnel high school and college graduates into these cities (compared to county cities).

We note we tried a variety of other potential amenity measures. Distances to a major highway and to navigable rivers have no effects, once market potential is controlled for. Kilometres 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. UNDERSIZED 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 more modest productivity losses.

The results have policy implications for China and we turn to these now. The basic conclusion is that migration restrictions have resulted in many undersized cities and the costs of being significantly undersized are high. As discussed above, the results in Tables 2 and 4 can be used to calculate a peak point for each prefecture city in China where net output per worker is maximized and then calculate a 95% confidence interval on that peak size. In 1997, based on column (1) Table 4 estimates, 51% of the 205 cities in the sample are significantly undersized, or to the left of the lower confidence limit. For column (2), Table 4 estimates, 62% of cities are significantly undersized, while for Table 2 estimates, 63% are significantly undersized. Note for about 20% of cities with high MS values, the lower confidence limit is non-positive, so none of these cities can be classified as undersized. Undersized cities are those with more typical MS values around 1·4, which are generally far below the lower confidence limit for all three sets of results. In summary, migration restrictions in China, which have constrained the growth of cities appear to have had severe effects. To be balanced, we note the results do suggest also that a few cities are significantly oversized, although there are not many cities in the relevant size ranges on which to base estimates. For Table 2 and Table 4 columns (1) and (2) estimates, respectively, 1%, 6%, and 3% of cities, presumably highly favoured cities, are significantly oversized.

Table 7 shows welfare losses for cities based on equation (14) ranked by percentile losses. Although almost all 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 6 and Figure 1. At the 50th percentile, that city's loss is 17%, in terms of net output per worker. Overall, the average (unweighted) loss is 30%. However, for 25% of cities, we are talking about losses over 28%, and for 10% of cities, losses over 69%. Allowing migration to these cities, as is now starting to happen, will allow them to operate much more efficiently. But that, of course, is only the tip of the iceberg. The gains to migrants relative to their current wages in the rural sector would be enormous.

TABLE 7

Per cent losses in net output per worker from operating away from the peak

Percentiles of cities (ranked by loss)	Largest loss (%)
First	
5% of cities	0·16
10%	0·76
25	3·8
50	17
75	38
90	69
95	103
100	229

One can imagine many caveats for this exercise. Foremost is that the recommendation here is not to suddenly increase the sizes of all cities by enormous magnitudes overnight. Underlying the process is adjustment in city management and construction of infrastructure that is buried in the formulation. The recommendation is to free up migration where migration responses take time as it is, giving cities room to adjust.

APPENDIX A. DERIVATION OF TEXT EQUATIONS.

To derive the equations in the text we first examine the maximization problem of producers and market clearing conditions.

A.1. *Firm profit maximization and entry*

Final producers. A final output producer seeks to maximize profits:

$$p_y AL^{\varepsilon} k_y^{\alpha} \ell_y^{\beta} \left(\int_{s_x} x(i)^{\rho} di \right)^{\gamma/\rho} - c_y - \int_{s_x} p_x(i) x(i) di - w\ell_y - rk_y. \tag{A0}$$

p_y is the price of the final good variety of a representative producer; w is the local wage rate; r is the fixed cost of capital in national or international markets; and $p_x(i)$ is the local price of intermediate input variety $x(i)$. Substituting in p_y from equation (5) in the text, maximization of (A0) yields first-order conditions:

$$\text{MP}^{1/\sigma_y}(y - c_y)^{-1/\sigma_y} \left(\frac{\sigma_y - 1}{\sigma_y} \right) \beta y/\ell_y = w \tag{A1}$$

$$\text{MP}^{1/\sigma_y}(y - c_y)^{-1/\sigma_y} \left(\frac{\sigma_y - 1}{\sigma_y} \right) \alpha y/k_y = r \tag{A2}$$

$$\text{MP}^{1/\sigma_y}\ (y - c_y)^{-1/\sigma_y} \left(\frac{\sigma_y - 1}{\sigma_y} \right) \gamma y/(s_x x) = p_x. \tag{A3}$$

The last condition (A3) for a single input variety, after differentiation then anticipates intermediate input symmetry where y producers each purchase x of any variety and buy s_x varieties. Substituting (A1)–(A3) into the profit function in (A0) set equal to 0 yields the equilibrium output for a single y producer, where gross output is

$$y = \sigma_y c_y. \tag{A4}$$

Intermediate good producers. For intermediate good producers, where $\eta_x = -(1-\rho)^{-1}$ and labour input usage is given (2), profit maximization gives the classic Dixit–Stiglitz results:

$$p_x = \frac{wc_x}{\rho} \tag{A5}$$

$$X = \frac{f_x \rho}{(1-\rho)c_x} \tag{A6}$$

$$\ell_x = \frac{f_x}{1-\rho}. \tag{A7}$$

A.2. *Local market clearing*

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

$$s_x \ell_x + s_y \ell_y = L \tag{A8}$$

$$X = s_y x \tag{A9}$$

(A8) is a full employment equation for s_x producers of x in the city and s_y producers of the traded good. (A9) states that supply of any variety, X, equals demand, where s_y producers each buy x of the intermediate input.

A.3. *Solving for employment allocations and numbers of firms*

First, we solve for the use of ℓ_y by y producers. Into (A1), substitute (A5) for

$$\ell_y = \frac{\beta}{\gamma} \frac{f_x}{1-\rho} s_x / s_y. \tag{A10}$$

Then by using (A8) we can get (A11), where

$$s_x = \frac{\gamma}{\gamma+\beta} \frac{(1-\rho)}{f_x} L. \tag{A11}$$

To solve for s_y, into the production relationship in (1), we substitute for k_y from (A2), ℓ_y from (A10), s_x from (A11), x from (A9), X from (A6), and y from (A4). The result is

$$s_y = Q_0^{1/(1-\alpha)} \mathrm{MP}^{\alpha/(\sigma_y(1-\alpha))} r^{-\alpha/(1-\alpha)} A^{1/(1-\alpha)} L^{(\varepsilon+\gamma/\rho+\beta)/(1-\alpha)}. \tag{A12}$$

Equations (A11) and (A12) give the number of intermediate and final good producers in a city, where the latter is an increasing function of city effective labour force and market demand/potential and a decreasing function of capital costs.

A.4. *Text equations*

(i) Equation (8).
Net output in the city is $(p_y(y-c_y) - rk_y)s_y$, which after substituting in equation (5) is $(\mathrm{MP}^{1/\sigma_y} (y-c_y)^{(\sigma_y-1)/\sigma_y} - rk_y)s_y$. Into this, substitute for rk_y from (A2) and for y from (A4) to get $\mathrm{MP}^{1/\sigma_y} Q_1 s_y$. From (A12) for s_y we then have (8).

(ii) Equation (10).
Into the total city value-added, $p_y(y-c_y)\, s_y$, which given (5) equals $\mathrm{MP}^{1/\sigma_y}(y-c_y)^{(\sigma_y-1)/\sigma_y} s_y$, substitute for $s_y$from (A12). This expression contains r, while we need an expression in K. Using (A2), (A12), and $k_y = K/s_y$, we solve for r in terms of K. Substituting in the revised expression for value-added gives (10).

(iii) Equation (11).
Value-added in the y sector is $p_y(y-c_y)s_y - p_x s_x x$ and in the x sector is $p_x s_x X$. Utilizing (A3), (5), and (A4) in the ratio of the two value-added expressions yields $\mathrm{MS} = (1-\gamma)/\gamma$ and equation (11).

APPENDIX B. DATA SOURCES

City-level data used in our analysis come from several sources. Most economic and amenity variables were taken from the 1991–1998 annual volumes (for data years 1990–1997) of the *Urban Statistical Yearbook of China* (hereafter *Yearbook*),[11] and *Cities China 1949–1998*. The latter includes a compilation of selected data in 1990–1997 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.* 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 U.S., or the

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

TABLE B1

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[14]	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 metres	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 centimetres from map S4	S3, S4
Distance to provincial capital	Horizontal distance from capital of province in which a city is located, measured in centimetres 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

Source: 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 edn., 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.

urbanized portion of a metropolitan statistical area. For 1997, we then start with a base sample of 223 prefecture-level cities for which we have labour force data (out of 226 official prefecture-level cities). For 217 of these we also have data for 1990 on labour force. We then exclude three oil-dominant cities,[12] and one city with unreliable data,[13] based on extraordinary year-to-year changes in labour force, which is likely the result of miscoding. The estimating sample is 205, where the other eight excluded cities have missing observations on variables used in either 1990 or 1997. Brief descriptions of the variables used in our analysis are in Table B1. We note capital is original book value of capital of industrial enterprises with independent accounting systems. For comparison of real growth of output (GDP), we use the provincial-level urban resident consumer price index to deflate nominal GDPs. 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).

[12] Daqing, Dongying, and Karamay. These cities are extreme outliers in terms of capital usage, with a second to third sector ratio in excess of 10.

[13] Jining of Shandong province.

[14] 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 B2

Urban variable means and S.D.

	Mean	S.D.
Output per worker	23,191	11,260
Capital per worker	30,579	18,318
Employment (in ten thousands)	53	67
% High school	22·5	8·26
Manufacturing to service ratio (GDP)	1·44	0·700
Accumulated FDI (in $) per worker since 1990	954	1582
Market potential	1452	375

Acknowledgements. 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. The paper has benefited greatly from comments by the editor, Bernard Salanié, and two anonymous referees, as well as seminar presentations at the Philadelphia Federal Reserve, LSE, Catholic University of Louvain, Bank of Italy, and University of Bologna.

REFERENCES

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

AU, C. C. and HENDERSON, J. V. (2005), "How Migration Restrictions Limit Agglomeration and Productivity in China", *Journal of Development Economics* (http://www.econ.brown.edu/faculty/henderson/papers/China402.pdf) (forthcoming).

BERGSMAN, J., GREENSTON, P. and HEALY, R. (1972), "The Agglomeration Process in Urban Growth", *Urban Studies*, **9**, 263–288.

BLACK, D. and HENDERSON, J. V. (1999), "A Theory of Urban Growth", *Journal of Political Economy*, **107**, 252–284.

BLACK, D. and HENDERSON, V. (2003), "Urban Evolution in the USA", *Journal of Economic Geography*, **11**, 343–373.

BLUNDELL, R. and BOND, S. (1998), "GMM Estimation With Persistent Panel Data" (IFS Working Paper No. W99/4, University College London).

BOTTAZI, L. and PERI, G. (2003), "Innovation and Spillovers in Regions: Evidence from European Patent Data", *European Economic Review*, **47**, 687–710.

CAI, F. (2000) *Zongguo Liudong Renkou Wenti* [*The Mobile Population Problem in China*] (Zhengzhou: Henan People's Publishing House).

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

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

DAVIS, D. and WEINSTEIN, D. (2001), "Market Size, Linkages, and Productivity: A Study of Japanese Regions" (NBER WP No. 8518, National Bureau of Economic Research).

DIXIT, A. and STIGLITZ, J. (1977), "Monopolistic Competition and Optimum Product Diversity", *American Economic Review*, **67**, 297–308.

DURANTON, G. and PUGA, D. (2001), "Nursery Cities", *American Economic Review*, **91**, 1454–1463.

DURANTON, G. and PUGA, D. (2004), "Micro-Foundations of Urban Agglomeration Economies", in J. V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics*, Vol. 4 (Amsterdam: North-Holland) ch. 48, 2063–2117.

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

FUJITA, M., KRUGMAN, P. and VENABLES, A. J. (1999) *The Spatial Economy* (Cambridge, MA: MIT Press).

FUJITA, M. and OGAWA, H. (1982), "Multiple Equilibria and Structural Transition of Non-Monocentric Configurations", *Regional Science and Urban Economics*, **21**, 161–196.

HANSON, G. (2005), "Market Potential, Increasing Returns, and Geographic Concentration", *International Economic Review*, **67**, 1–24.

HEAD, K. and MAYER, T. (2004), "The Empirics of Agglomeration and Trade", in J. V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics*, Vol. 4 (Amsterdam: North-Holland) ch. 59, 2609–2669.

HELSLEY, R. and STRANGE, W. (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*, **64**, 640–656.

HENDERSON, J. V. (1988) *Urban Development: Theory, Fact and Illusion* (New York: Oxford University Press).

HENDERSON, J. V. and WANG, H. G. (2005), "Urbanization and City Growth" (Brown University; http://www.econ.brown.edu/faculty/henderson/papers/UrbanizationandCityGrowth0405.pdf).

HUMMELS, D. (2004), "Towards a Geography of Trade Costs" (Mimeo, Purdue University).

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

KOLKO, J. (1999), "Can I Get Some Service Here: Information Technology, Service Industries, and the Future of Cities" (Mimeo, Harvard University).

LIN, J. Y., CAI, F. and LI, Z. (1996) *The China Miracle: Development Strategy and Economic Reform* (Hong Kong: 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", in J. V. Henderson and J.-F. Thisse (eds.) *Handbook of Urban and Regional Economics*, Vol. 4 (Amsterdam: North-Holland) ch. 59, 2243–2291.

MUESER, P. and GRAVES, P. (1995), "Examining the Role of Economic Opportunity and Amenities in Explaining Population Redistribution", *Journal of Urban Economics*, **37**, 176–200.

OVERMAN, H. G., REDDING, S. and VENABLES, A. J. (2003), "The Economic Geography of Trade, Production, and Income: A Survey of Empirics", in J. Harrington and K. Choi (eds.) *LSE Handbook of International Trade* (Malden, MA: Basil Blackwell) 353–387.

PERKINS, D. (1994), "Completing China's Move to the Market", *Journal of Economic Perspectives*, **8**, 23–46.

PONCET, S. (2005), "A Fragmented China: Measure and Determinants of Chinese Market Disintegration", *Review of International Economics*, **13**, 409–430.

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

ROSENTHAL, S. and STRANGE, W. (2004), "Evidence on the Nature and Sources of Agglomeration Economies", in J. V. Henderson and J.-F. Thisse (eds.) *Handbook of Urban and Regional Economics*, Vol. 4 (Amsterdam: North-Holland) ch. 49, 2119–2171.

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.

TOLLEY, G., GARDNER, J. and GRAVES, P. (1979) *Urban Growth Policy in a Market Economy* (New York: Academic Press).