Queen’s Economics Department Working Paper No. 75

THE SIZES AND TYPES OF CITIES

J.V. Henderson
Queen’s University

Department of Economics
Queen’s University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

11-1972

THE SIZES AND TYPES OF CITIES*

J. V. Henderson

Discussion Paper No. 75

Queen's University
(Revised November 1972)

*The author is indebted to Charles Upton, George Tolley, and Harry Johnson for detailed comments on the analytical content and organization of earlier versions of the paper.

This paper presents a simple general equilibrium model of an economy where production and consumption occur in cities. The paper focuses on the different sizes and types of cities generated by market forces and whether these market forces generate optimum size cities. Before the model is presented, four complex questions are naively answered, revealing the most basic concepts underlying the paper and intellectual debts to the existing literature.

Why does an economy have cities, in particular large cities, instead of production and consumption being organized in homesteads spread out across the countryside? We have population agglomeration or cities because there are technological economies of scale in production and/or consumption and because these activities are not space or land intensive (relative to, say, agriculture). Scale economies may occur at the final output level, at the marketing level, or at the intermediate input level such as in transportation, natural resource extraction, and capital and labour market development.

What limits city size? Hypothesizing that agglomeration occurs due to scale economies in production of a city's traded good, Mills [13] demonstrates that activity associated with traded good production involves increasing per unit resource costs. In particular, workers employed in traded good production must be housed. Mills assumes traded good production occurs in the Central Business District (C.B.D.) and housing is located around the C.B.D. People commute from their homes to work and back daily. As city size and the area devoted to housing increase spatially, the average distance a worker commutes and congestion necessarily increase. Therefore the per person resource costs of commuting in terms of workers' time and expenditure on transportation facilities increase with city size. Efficient city size is achieved when these increasing per person resource costs offset the initial resource savings due to scale economy exploitation in traded good production. This type of argument plays a crucial role in this paper.

Lösch [11] and Beckman [2] argue city sizes are limited by the extent of the market for their products. The determination of the market area of a city involves trading off the benefits of further scale econony exploitation with an increase in city size against the increased transportation costs of having to sell the additional city production in markets farther from the city. This proposition is discussed at the end of the paper.

Why do city sizes vary? City sizes may vary because of differing efficiency of city organization and public good provision and differing access to export and input markets. Even without these differences, city sizes would differ. Different types of cities exist specializing in the production of different traded goods. If these traded goods involve varying degrees of scale economies, the average amount of increasing per person commuting costs these cities can support varies under the Mills' model. Hence city size varies. But why do cities specialize?

Providing there are no positive production benefits from locating two industries together (such as using a common labour market), locating the production of the two goods in the same city only works to raise total production costs. Labourers employed in the two industries both contribute to rising per person commuting costs, but scale economy exploitation occurs only with labour employment within each industry. If we locate the industries together, there are higher average per person commuting resource costs for a given level of scale economy exploitation or industry employment within either industry, than if we locate the industries in separate cities. This is one reason why cities will tend to specialize in the production of different traded goods. To be weighed against the specialization argument are the transportation costs of executing trade between two or more specialized cities. Goods such as retailing services are non-traded between cities because of high transportation costs.

Note that cities will tend to specialize in bundles of goods. Within each bundle of goods, the goods will utilize a common labour market such as for electronics experts in television, radio, typewriters, photocopying, or for industries employing separately the male and female workers from a household. Cities could have interconnected industries such as the tourist industry and its facilities and the convention and specialized business and sales training industries. Alternatively, within a bundle of goods, the goods could use a common intermediate input that is difficult to transport. Examples range from private industry intermediate inputs to public goods such as intra-city transportation systems attracting warehousing and trucking industries.

The final question is whether the size different types of cities attain in a market economy is optimal? City size may be non-optimal because of inefficient pricing of congestion in commuting or of the output of goods that are produced with economies of scale external to the firm or because of other externalities such as pollution. While paying tribute to these problems, this paper will focus on another set of problems. These problems concern the market forces that generate cities, especially the market forces caused by the behaviour of capital owners, labourers, and firms.

In the paper, first the model of a single city is presented. How factor rewards and cost of living vary with city size is analyzed. Given these results, the paper presents an analysis of market equilibrium and optimum city size. Finally equilibrium in an economy with multiple types of cities is examined. At the end of the paper we discuss how natural resources and transportation costs in trade can be integrated into the model.

Throughout the paper, it is assumed capital and labour are scarce resources and perfectly mobile within the economy. The economy is situated on a flat featureless plain, large enough so land per se is never a scare resource (although location will be a scare factor). This non-critical assumption implies the opportunity cost of land is zero. There are no specified transport costs of inter-city trade.

1. THE MODEL OF A CITY

The model of a single city is presented in this section and solved for factor reward equations. Factor rewards will be a function of the quantity of capital and labour employed in the city and the price of the city's export good. In the next section, factor reward equations are analyzed to solve for equilibrium and optimum city size. For example, when solving for optimum city size, the goal will be to maximize deflated factor rewards which will reflect a maximization of the consumption bundle available to the city and economy.

Although the model presented here is very different from the Mills [13] model, its prototype is the Mills' model. The city produces an export good in the Central Business District (C.B.D.) and housing in the area surrounding the C.B.D. Workers commute daily to and from the C.B.D. It is hypothesized that non-C.B.D. employment in residential shopping centers and suburban manufacturing plants is a secondary feature of urban organization. This employment is closely linked to the spatial and economic structure of the core city affecting and being affected by its rent pattern and economic characteristics. Economic activity in de-centralized areas of the city is intertwined with activity in the C.B.D., particularly with respect to the range or bundle of goods the city produces. When suburbs become virtually independent with little cross commuting to the core city and with weakened interdependence in the primary and intermediate input structure of the manufacturing and service industries, they become economic cities in their own right.

Production Conditions

The city produces a traded good, X_1, under conditions of increasing returns to scale. These increasing returns to scale are hypothesized to be internal to the industry in the city but external to the firm. The economies of scale could be due to economies in developing the city labour and capital markets or due to economies in the industry from utilizing an intermediate input not specified separately in the production function.

The industry production function is

$$X_1^{1-\rho_1} = L_1^{\alpha_1} K_1^{\beta_1} N_1^{\zeta_1}, \quad \alpha_1 + \beta_1 + \zeta_1 = 1, \quad 0 < \rho_1 < 1 \qquad (1)$$

where L_1, N_1, and K_1 are inputs of home or land sites, labour, and capital respectively. ρ_1 represents the degree of increasing returns to scale (i.e. $(\alpha_1 + \beta_1 + \zeta_1)/1 - \rho_1 > 1$) and the reason for production agglomeration at a point in space. It is assumed for now that X_1 is sold by the city at a fixed price, q_1, set in national or international markets.

Because economies of scale are <u>external</u> to the firm, the firm views itself as having a constant returns to scale production function (an economically non-critical assumption) and calculates marginal products accordingly. For example, the marginal product of labour as seen by the firm is $\zeta_1 \, x_1/n_1$ rather than the social marginal product $\frac{\zeta_1}{1-\rho_1} x_1/n_1$ where lower case X_1 and N_1 refer to firm output. This divergence between private and social marginal products preserves exhaustion of factor payments.[1] Perfect competition is

[1] The points in this paragraph are documented in Chipman [5] and Herberg and Kemp [7].

preserved because there are no economic barriers to entry; any entering firm immediately benefits from the industry scale economies. However the divergence between social and private marginal costs prevents the exact attainment of a Pareto-optimum allocation of factors. This problem will be discussed later in the paper, but does not play a major role in the analysis.

Within the city, housing services must be produced for labourers employed in the production of X_1. Housing services are an inter-city non-traded good whose prices may consequently vary between cities of different sizes. This price variation will account for the variation in the cost of living between cities of different sizes in the model. Housing is produced with constant returns to scale and inputs of capital, labour, and homesites.

$$X_3 = N_3^{\zeta_3} K^{\beta_3} L_3^{\alpha_3}, \quad \zeta_3 + \beta_3 + \alpha_3 = 1 \tag{2}$$

In a spatial model, a homesite represents an input of raw land and a spatial location in the city. If each city employs only one labourer utilizing raw land to produce homesites is costless. (The opportunity cost of raw land was previously assumed to be zero.) As city size grows beyond the homestead, in the first ring of housing around the production site, people spend, say, one unit of time commuting to utilize their homesite. In the second ring they must spend, say, two units of time to utilize the same size homesite. As city size continues to expand people must spend more and more time commuting to utilize their raw land and thus the average resource costs of producing homesites or housing <u>location</u> increases with city size. This increasing resource cost acts to offset the benefits of agglomeration in the production of X_1.

Without crucial omissions in economic reasoning, for algebraic simplicity, the spatial world is collapsed into a non-spatial world in this paper.

It is hypothesized that the resulting model works "as if" it were a spatial model; our results would be duplicated in a spatial model. Primarily the concern is to capture the increasing average resource costs of commuting that arise with increased city size.[2] To do this, an input into housing called homesites is specified.

Homesites or locations are produced with labour or commuting time inputs; raw land is not specified in the production function since its opportunity cost is zero.

$$L^{1-z} = N_0, \qquad -1 < z < 0 \tag{3}$$

where L and N_0 are homesite output and labour inputs respectively and z represents the degree of decreasing returns to scale implying rising average resource costs of increased homesite production and city size. Furthermore the degree of decreasing returns to scale, z, is assumed to increase with city size. Specifically it is assumed

$$1/1-z = N_A^{m}, \qquad -1 < m < 0 \tag{4}$$

where N_A is city population. The reason for this assumption is algebraic and will become apparent later in the paper.

[2] The other crucial aspect of the commuting phenomenon in a spatial model is land rents. Residential location theory as in Muth [16], Alonso [1], and Mills [13] tells us there is a spectrum of commuting costs and land rents in a city. The land rents act as an ordering device so that people who live nearer the C.B.D. and experience lower commuting costs pay higher rents to offset their cost advantage relative to those further from the C.B.D. The actual land itself involves no resource costs if its opportunity cost is zero. The land rents are a transfer from renter to landowner reflecting the relative "scarcity" of a location. As city size grows, land rents rise with commuting costs.

In a non-spatial there is no role for an ordering device or spectrum of land rents and landowners. Rising resource costs of commuting are captured but the location "scarcity" principle is not represented. However, given land rents are essentially a transfer from renter to landowner, our results concerning equilibrium and Pareto-optimum city size are unaffected. But to the extent that rising land rents induce further substitution away from homesite inputs in housing and X_1 production, the resource costs of the commuting phenomenon are "underrepresented" in our model.

Note that in this spaceless model, every person buys equal amounts of homesites at the same average price which rises with city size because of decreasing returns to scale in homesite production. The price rise indicates the rising resource costs of city size to be compared with the benefits of agglomeration in X_1 production.

Parallel to the production specification of X_1, the production of L is assumed to occur with (dis)economies of scale external to the firm or individual but internal to the industry. Intuitively, this says when an additional person moves to the city, he imposes unpriced externalities on the inhabitants of the city by raising their average commuting time as city size expands.[3] Because of the externality L is priced at private not social marginal cost and labour is paid the value of its private not social marginal product, implying, as for X_1, that factor payments are exhausted but output is not Pareto-optimal.

The final equations on the production side of the model are the resource and intermediate input employment equations where

$$\begin{aligned} N_1 + N_3 + N_0 &= N_A \\ K_1 + K_3 &= K_A \end{aligned} \qquad (5)$$

$$L_1 + L_3 = L \qquad (6)$$

where N_A and K_A are the labour force and capital stock of the city.[4]

[3] This externality formulation may not be strictly correct in a spatial model. For example, if lot size is fixed and there is no congestion, when an additional person moves to the city edge, he affects no one else's commuting costs and imposes no externality. If there is congestion or lot size is variable, the new inhabitant theoretically affects all commuting costs and lot sizes. See Muth [16], Chapters 2 and 3 on this point.

[4] This is the appropriate place to introduce two equations that will be used in footnotes later in the paper. Assuming firms through cost minimization pay factors the value of their private marginal products, it is

Consumption Conditions

To close the model, consumption conditions must be specified in order to derive demand equations for the three consumer goods. In addition to X_1 and X_3, city inhabitants purchase an import good X_2, at fixed price q_2.

It is assumed that consumers have identical tastes and logarithmic linear utility functions. These two assumptions allow us to maximize either every individual's utility function or one utility function for the whole society and obtain identical equilibrium conditions. Utility is maximized subject to income, Y, and output prices where q_1 and q_2 are fixed and q_3 is free to vary with city size. Y is determined as follows.

In the case of labour income, labourers live in the city where they work and therefore it is assumed that all labour income earned in the city is spent in the city. Capital owners are not constrained to live in the city where their capital rentals are earned. They may live in the countryside of our flat featureless plain, in other cities, or in other countries. Cities may be net borrowers or net lenders with respect to the proportion of capital rentals earned versus spent in the city. Given this problem, two alternative _polar_ assumptions are made.

Assumption A. All capital owners live in the cities of _this_ economy and also work as labourers. Since capital ownership could be unevenly distributed amongst labourers and cities, any one city could be a net borrower

4. Continued.
well known the following indirect production and cost functions may be derived.

$$X_1 = \delta_1^{-\delta_1/\rho_1} \alpha_1^{-\alpha_1/\rho_1} \beta_1^{-\beta_1/\rho_1} q_1^{1/\rho_1} p_N^{-\delta_1/\rho_1} p_L^{-\alpha_1/\rho_1} p_K^{-\beta_1/\rho_1}$$

$$q_3 = \alpha_3^{-\alpha_3} \beta_3^{-\alpha_3} \delta_3^{-\delta_3} p_N^{\delta_3} p_L^{\alpha_3} p_K^{\beta_3}$$

or lender. To avoid problems of arbitrarily declaring some cities net lenders and others net borrowers as well as other problems discussed below, we assume capital ownership is evenly divided amongst labourers. Then the proportion of capital rentals earned to capital rentals spent in the city will be one (until section 3, we assume the capital-labour ratios of the city and country are the same). However it is not assumed that the capital owners who live in a city invest in that city. They can invest in other cities while capital owners in other cities invest in their city.

Assumption B. Capital owners are a separate group of people who do not work as labourers and who live in the countryside or other countries. No capital rentals are spent in the cities of this economy. Later we will see this assumption affects city size relative to assumption A, because less income and production resources are devoted to homesites in the city and housing and the prices of these goods to labourers rise more slowly with city size.

Given the real world is a mixture of these polar cases, they have been isolated because they imply polar differences in optimal investment behaviour of capital owners and in equilibrium and optimum city size. The arguments will pivot on the fact that, under assumption A, the welfare of capital owners who live in cities is influenced by the city cost of living. Also as mentioned above, the fact capital rentals are spent in the cities increases the relative amounts of resources devoted to homesite and housing production in cities and the cost of these. Neither of these facts pertain to assumption B.

To summarize our consumption conditions, for the city

$$U = X_1^a X_2^b X_3^c \qquad a+b+c = 1 \tag{7}$$

is maximized subject to, for assumption A and B respectively, either

$$Y = p_N N_A + p_K K_A \qquad \text{or} \qquad Y = cp_N N_A$$

From utility maximization, demand equations can be obtained such that

$$X_1^C = aY/q_1, \; X_2^C = bY/q_2, \; X_3^C = cY/q_3 \tag{8}$$

where the superscript C [P] refers to goods consumed [produced] in the city. Balance of payments equilibrium in trade requires exports equal imports plus net capital rental outflows or that

$$X_1^P q_1 - X_1^C q_1 = X_2 q_2 + kp_K K_A$$

where k is the proportion of capital rentals earned in the city leaving the city. k equals one or zero under assumption B or A respectively.[5]

By substituting equations (8) into (7) the indirect utility function

$$U = a^a b^b c^c y \; q_1^{-a} \; q_2^{-b} \; q_3^{-c} \tag{9}$$

is obtained where y is the income of the person whose utility is being examined. This formulation will be utilized throughout the paper.

2. SOLUTION OF THE MODEL FOR A CITY

From the consumption and production equations of the model city output, factor payments, and the price of homesite and housing can be solved for in terms of city employment of capital (K_A) and labour (N_A), and prices of traded goods, q_1 and q_2 which are fixed in this section. The solutions are obtained by combining the value of private marginal product equations and consumer demand equations obtained by normal optimization behaviour with the full employment equations. The actual method of solution may be found

[5] Equation (8) for X after substituting in the different values for Y becomes under (a) assumption A: $X_3 = c/1\text{-}c \; q_1 X_1^P$ and (b) assumption B: $X_3 = cp_N \; N_A/q_3$.

in Henderson (1972) or pieced together from the footnote below.[6]

Here the appropriate expressions for factor rewards in terms of K_A and N_A are presented and in the next section we show how factor rewards vary with city size. The variation in factor rewards will lead to the conclusion that there is an efficient size of city; i.e., one that maximizes factor rewards. First we must discover what the appropriate measures of factor rewards are -- measures reflecting the welfare of factor owners, which we wish to maximize.

For labourers the appropriate measure of factor rewards is some index of utility. Since, by assumption, labourers live in cities, their utility is affected by the wage rate and the city cost of living or housing, as indicated by equation (9). Not only does equation (9) give a measure of the welfare of labourers, but it indicates their behavior in choosing a city to live in, since they will choose the city that maximizes their welfare. In discussing both market equilibrium and optimum city size, changes in utility with city size will be examined. The different values that utility assumes with changing city size will form a utility path.

[6] The steps to attain our solutions are as follows. First we multiple the factor or intermediate resource supply equations, (5) and (6), by p_N, p_K, and p_L respectively. We then substitute in expressions for their components from the value of factor marginal product equations; e.g., from (1) and (3), $p_N N_1 = q_1 \zeta_1 X_1$ or $p_N N_0 = p_L L$. In the resulting three equations for X_1, $p_L L$ and $q_3 X_3$, we substitute in the expression in footnote 4 and then equation (3) for N_0, and the expression in footnote 5, respectively. We then have the three original resource supply equations in terms of p_N, p_L, p_K, N_A and K_A. By adding the equations originally derived from labour and homesite supply, we get p_L in terms of p_N, p_K, and N_A. We substitute this expression for p_L into the equation originally from capital supply and solve for p_K. We take these equations for p_L and p_K and substitute them back into the equation originally from labour supply. Our solutions for p_N, p_K, and p_L may then be written out. We can use these to get expressions for q_3 and X_1 (see footnote 4) as well as utility (see equation (9)).

For capital owners the distinction between variables governing market or investment behaviour and the benefits or utility from such behaviour is crucial. When making investment decisions, capital owners seek the highest capital rental possible, regardless of where they live. Since they are not constrained to live in the city where their capital is employed or in any city at all, the price of housing or cost of living in a particular city does not influence their investment decisions.[7] In terms of spending their capital receipts, if capital owners live outside cities in the countryside or other countries as under assumption B, the city cost of living index does not affect their welfare and they are best off maximizing capital rentals. That is, they maximize the income variable entering their utility function in equation (9), where the consumer prices (in particular housing) they pay are independent of their investment returns in the cities of the economy.

If assumption A prevails and capital owners live in the cities of this economy, the benefits of spending their capital receipts are affected by the city cost of living. Since, as we will see, the capital rentals they earn and the price they pay for housing are directly related to and affected by city size, the ultimate consumer benefits they get from investing are indicated by a utility index that includes capital rentals and housing prices. Note for assumption A although we have assumed capital owners are also labourers, we have dichotomized their income and actions by income source.

[7] If one assumes capital is physically tied to the owner as for, say, a small business owner, then he will move his business to maximize an index of utility--the capital rentals deflated by a cost of living. For example a small business owner in New York will demand a higher return on his capital than if he lived in Albany, simply due to cost of living differences.

Under our current assumptions this separation is technically valid, at least until section 3.[8]

Given the above discussion, the following equations for factor rewards are given and will be used in the subsequent analysis of city size. (Their derivation is described in footnote 6.) The exponents of the equations contain production and consumption parameters including $1/1\text{-}z = N_A^m$, $0 > m > -1$, the degree of decreasing returns to scale in homesite production. The constants contain similar coefficients. The variables in the equations are total city employment of capital and labour.

Assumption A. The location decision of labourers and their welfare levels are reflected in the solution for the utility of a labourer (recall all labourers receive identical income and have identical tastes).

$$U_N = \left[W_N q_2^{-b} q_1^{1-c-a} {}_t \; N_A^m \left| \frac{\alpha_1(1-c) + c\alpha_3(1-\rho_1)}{\rho_1 - 1} \right| \right] \cdot$$

$$\left[(K_A/N_A) \; \frac{-\beta_1 - c\beta_3 + c\beta_1 + c\beta_3\rho_1}{\rho_1 - 1} \right] \cdot$$

$$\left[N_A \; \frac{(\alpha_1(1-c) + c\alpha_3(1-\rho_1))(1-N_A^m) - \rho_1(1-c)}{\rho_1 - 1} \right] \tag{10}$$

[8] We assumed capital ownership is equally divided amongst labourers. This assumption allows us to circumvent the effect of capital earnings upon labour location decisions. We hypothesized labourers move to equalize p_N deflated by q_3 *between* cities. Labourers will actually move to equalize p_N + their share of $\overline{p_K}K$, both deflated by q_3, the city cost of living. $\overline{p_K}$ is *exogenous* to location decisions since it is determined by investment behavior where p_K's are equalized between *all* cities. If capital ownership is evenly divided amongst labourers and if all cities are of the same size and type (as will be the case until section 3), q_3 will be equal between cities. Thus it will also be true that $p_N + \overline{p_K}\; K_A/N_A$ deflated by q_3 will be equal between cities when p_N deflated by q_3 is. However if a city is a net lender its capital earnings will be higher than that of a net borrower, which we will see implies q_3 or the city cost of living is higher. If a net lender has a higher q_3, for p_N plus a share of $\overline{p_K}K$ deflated by q_3 to be equal between the cities, p_N deflated by q_3 must be relatively higher in the net lending city. In the discussion to follow this would lead to minor variations in p_N's

where W_N and t^9 are constants and $\partial U_N/\partial(K_A/N_A) > 0$.

The investment decisions of capital owners are governed by their capital receipts where

$$P_K = C_K q_1 t^{\frac{\alpha_1 N_A^m}{\rho_1 - 1}} (K_A/N_A)^{\frac{1-\rho_1-\beta_1}{\rho_1-1}} N_A^{\frac{\alpha_1(1-N_A^m)-\rho_1}{\rho_1-1}} \tag{11}$$

where C_K and t are constants and $\partial p_K/\partial(K_A/N_A) < 0$ if $\rho_1 + \beta_1 < 1$.

The benefits to capital owners of spending their capital receipts is given by the expression for utility

$$U_K = \left[W_K q_2^{-b} q_1^{1-c-a} t^{N_A^m\left[\frac{\alpha_1(1-c)+c\alpha_3(1-\rho_1)}{\rho_1-1}\right]} (K_A/N_A)^{\frac{1-\rho_1-\beta_1+c\beta_1+c\beta_3(\rho_1-1)}{\rho_1-1}}\right] \cdot$$

$$\left[N_A^{\frac{(\alpha_1(1-c)+\alpha_3 c(1-\rho_1))(1-N_A^m)-\rho_1(1-c)}{\rho_1-1}}\right] \tag{12}$$

where W_K and t are constants and $\partial U_K/\partial(K_A/N_A) < 0$, if the exponent of (K_A/N_A) is positive.

<u>Assumption B</u>. Under assumption B, as discussed above, we are only concerned with the expressions for capital rentals and the utility of labourers. These two variables indicate both what factor owners seek to maximize in the market and the consumption benefits from their actions. The expressions for U_N and p_K under assumption B are <u>identical</u> to those under assumption A except the constants W_N, C_K, and t are replaced by W_N^1, C_K^1 and s^{-1}[10] respectively.

deflated by q_3's, in K_A/N_A ratios of cities, and city sizes. Our assumption avoids these quantitative complications.

[9] $t = (\alpha_1 + \alpha_3\frac{c}{1-c})^{-1}(\alpha_1 + \delta_1 + (\alpha_3 + \delta_3)\frac{c}{1-c}) > 1$. The values of constants C_j and W_j are in Henderson (1972). They are not needed here.

[10] $s^{-1} = [(1 - \delta_3 c - \alpha_3 c)\alpha_1(\alpha_1 + \delta_1)^{-1} - \alpha_3 c]^{-1} > 1$.

The equations are not repeated here.

The equations presented above are a function of the K_A/N_A ratio in the city, a measure of scale of city output or N_A, and a variety of production and consumption parameters represented in the exponents and constants of the equations. The normal general equilibrium factor ratio effects upon factor rewards are present in that, $\partial p_K/\partial(K_A/N_A) < 0$, $\partial U_N/\partial(K_A/N_A) > 0$, and $\partial U_K/\partial(K_A/N_A) < 0$, unless the scale effect in traded good production, or ρ_1, is too large. In the following section we examine how factor rewards vary with city size and what the crucial production and consumption parameters are in determining this.

Utility and Capital Rental Paths

In discussing how factor rewards vary with city size, the scale effect upon factor rewards will be isolated from the K_A/N_A ratio effect and the effect of any changes in the prices of traded goods. To do this, the derivative of the various factor rewards is taken with respect to N_A or city size, holding the K_A/N_A ratio and traded good prices constant. Taking the changes in factor rewards, the values that factor rewards assume ~~with~~ city size will be summarized in factor reward paths. The effect, in terms of shifting these paths, of changes in the K_A/N_A ratio or output prices will also be examined.

Assumption A. Our factor reward paths are derived first under assumption A where capital owners live in cities. To determine the paths we take the derivative of the logarithm of equations (10), (11) and (12) with respect to N_A holding K_A/N_A fixed:

$$\frac{dp_K}{dN_A} = N_A^{m-1}\, p_K \frac{\alpha_1 m}{\rho_1 - 1} \left[\log t - 1/m + \frac{\alpha_1 - \rho_1}{\alpha_1 m} N_A^{-m} - \log N_A\right] \tag{13}$$

$$\frac{dU_K}{dN_A} = U_K N_A^{m-1} \left(\frac{m(\alpha_1 - c\alpha_1 - c\alpha_3(\rho_1 - 1))}{\rho_1 - 1}\right) \left[\log t - 1/m + \frac{(1-c)(\alpha_1 - \rho_1) + c\alpha_3(1-\rho_1)}{m(\alpha_1 - c\alpha_1 - c\alpha_3(\rho_1 - 1))} N_A^{-m} - \log N_A\right] \tag{14}$$

$$\frac{dU_N}{dN_A} = \frac{U_K}{U_N}\frac{dU_K}{dN_A} , \quad \text{for } \frac{dU_K}{dN_A} \text{ in equation (14)} \qquad (15)$$

Our results from analyzing equations (13)-(15) are summarized as follows:

(1) The sign of the derivatives are given by the sign of the expression in the square brackets in each equation.

(2) When N_A is small, the expressions in the square brackets and the derivatives are all positive, indicating that initially capital rentals and utility levels rise as city size rises.[11]

(3) As N_A increases, either the derivatives remain positive or become negative, depending on whether the signs of the third terms in the square brackets are positive or negative. If $\alpha_1 \geq \rho_1$, the derivatives will eventually become negative[12] indicating capital rentals and utility levels will rise to a maximum and then decline.[13] If $\alpha_1 < \rho_1$, the derivatives are always positive[14]; and capital rentals and utility levels rise indefinitely with city size.

[11] log $t > 1$; $1/m > 0$; and, for $N_A \to 1$, $-\log N_A \to 0$. Therefore, for $N_A \to 1$, dp_K/dN_A, $dU/dN_A > 0$.

[12] This is a necessary condition for dp_K/dN_A but only a sufficient condition for dU/dN_A. If $\alpha_1 < \rho_1$ and $|(1-c)(\alpha_1-\rho_1)| < c|c\alpha_3(1-\rho_1)$, U will also achieve a finite maximum. Since p_K rises indefinitely when $\alpha_1 < \rho_1$, this means there is a range of parametric values where U achieves a maximum and declines, while p_K rises indefinitely. Due to space limitation, in this paper (cf. Henderson [6]), we only examine the situations where either both U and p_K rise indefinitely or both U and p_K achieve a finite maximum and decline.

[13] At this point we mention the reason for having $N_A^m = 1/1-z$. For example, returning to equation (11), replacing N_A^m by $1/1-z$ where z is fixed, and taking the derivative of p_K, we find that $dp_K/dN_A > 0$ for all values of N_A (or less than zero for all values). The p_K path never achieves a maximum, which is the case of interest in the analysis to follow.

[14] Note that the $-\log N_A$ in the square brackets, for the case $\alpha_1 < \rho_1$, could result in a local maximum in the rising factor reward paths. This unlikely possibility is discussed in Henderson [6].

That factor rewards may not increase indefinitely arises because the benefits of agglomeration, represented by scale economies (ρ_1) in X_1, may be offset by the disadvantages of scale economies (z) in homesite production. Recall also that z increases with city size ($1/1-z = N_A^m$, $0 > m > -1$), insuring accelerating resource costs of homesite production. Whether factor rewards actually achieve a maximum and decline depends on the relative values of α_1 and ρ_1 in equations (13)-(15). While ρ_1 represents resource savings, α_1 represents the homesite intensity of traded good production--the derived demand for homesites, a good produced with rising per unit resource costs. Other things being equal, a rise in α_1 would increase the relative production requirement of the resource costly homesites. If α_1 is large enough so $\alpha_1 \geq \rho_1$, then eventually homesite resource costs will lead factor rewards to start to decline.[15]

(4) If the paths achieve a finite maximum and then decline the following points are germane. The maximum points, $N_A(p_K^*)$, $N_A(U_N^*)$, and $N_A(U_K^*)$, may be obtained by equating equations (13)-(15) to zero and solving for N_A. As is obvious from equations (14) and (15), $N_A(U_N^*) = N_A(U_K^*)$. But it can also be shown $N_A(U_N^*) = N_A(U_K^*) < N_A(p_K^*)$.[16] That $N_A(p_K^*) > N_A(U_N^*, U_K^*)$ is not surprising since U_N and U_K represent consumption benefits, not just production benefits as p_K does. From equation (9), we see that U_N or U_K are p_N or p_K

[15] Note that in a more complex model, ρ_1 as well as z could change with city size too. This would reduce the rigidity of the condition that $\alpha_1 \geq \rho_1$ overall ranges of production for p_K to have a finite maximum. If ρ_1 declined with city size, our condition $\alpha_1 \geq \rho_1$ might just be a marginal one.

[16] The expressions in the square brackets of equations (13) and (14) differ only by the coefficient of the third term. From these coefficients where equations (13)-(15) are equated to zero, if

$$\left|\frac{(1-c)(\alpha_1-\rho_1) + c\alpha_3(1-\rho_1)}{m(\alpha_1-c\alpha_1-c\alpha_3(\rho_1-1))}\right| > \left|\frac{\alpha_1 \; \rho_1}{\alpha_1 m}\right| \text{ then } N_A(U_N^*, U_K^*) < N_A(p_K^*).$$

deflated by q_3^{-c}, the price of housing. This fact is indicated by the c and α_3 parameters that appear in equations (14) and (15), representing the share of housing in consumption and of homesites in housing production. Both parameters represent derived demand for homesites, our resource costly good. The consumption effect of having to buy housing whose cost escalates with city size is to reduce U_K and U_N relative to p_K and p_N. Thus $N_A(U_N^*, U_K^*) < N_A(p_K^*)$.

(5) What happens to our capital rental and utility paths if K_A/N_A or q_1 change, particularly in the situation of major concern in this paper where $\alpha_1 \geq \rho_1$ and the paths achieve a finite maximum? From equations (10), (11), and (13) we saw if $\alpha_1 \geq \rho_1$, $\partial p_K/\partial(K_A/N_A)$, $\partial U_N/\partial(K_A/N_A)$, and $\partial U_K/\partial(K_A/N_A)$ and similar expressions for the partial derivative of q_1 are all positive. Therefore a fall in K_A/N_A or a rise in q_1 will shift the factor reward paths up at all points. Further, from equations (13)-(15) we can see that a change in K_A/N_A or q_1 does not affect the terms in the square brackets and hence does not change the point where the paths achieve a finite maximum.

Points (1) to (5) are illustrated in figure 1.

Assumption B. Having discussed factor reward paths under assumption A, the paths under assumption B where capital owners live outside the cities of the economy are examined. Only the U_N and p_K paths are of concern under assumption B, since p_K represents both investment returns and consumption benefits from such returns, as discussed above.

Recall that U_N and p_K were identical under assumptions A and B except for the values of constants. To examine how U_N and p_K change under assumption B, in equations (13) and (15) the only difference is that t is replaced by s^{-1} in the square brackets. The coefficients of the third terms in the square brackets. The coefficients of the third terms in the square brackets are

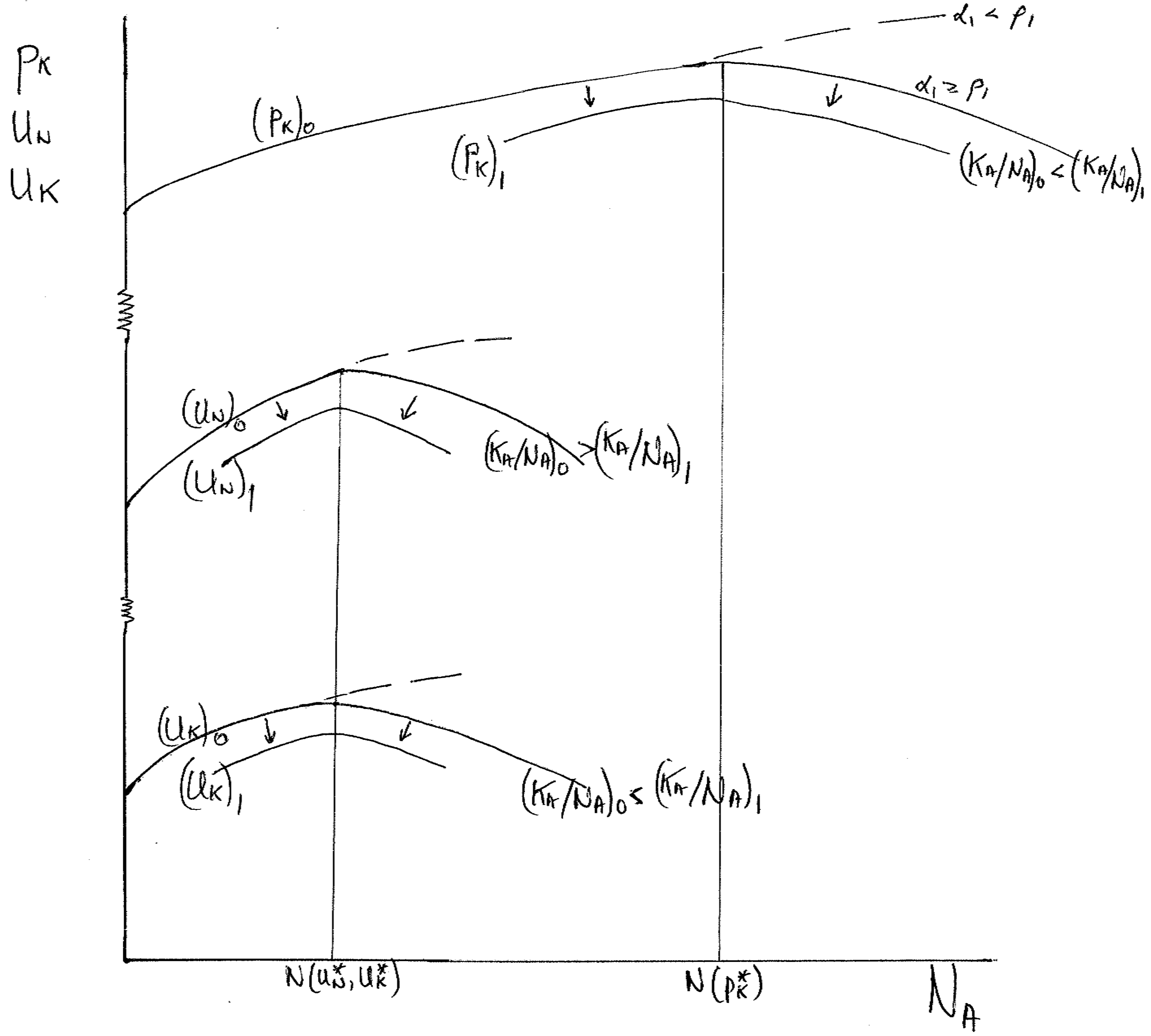

Figure 1: Capital Rental and Utility Paths under Assumption A

unchanged and thus the parametric conditions for whether the paths achieve a finite maximum are unchanged. In addition their relative positioning is unchanged and U_N achieves a finite maximum before p_K. Also $\partial U_N/\partial(K_A/N_A) > 0$, and $\partial p_K/\partial(K_A/N_A) < 0$, if $\alpha_1 \geq \rho_1$; and thus the paths shift as before with a change in K_A/N_A ratio.

However because s^{-1} and t are different, when the expressions in the square brackets of the new (13) and (15) are set equal to zero, different city sizes where U_N and p_K achieve a maximum result. Specifically because $-\log s > \log t$ the paths under assumption B achieve their maximum later than under assumption B.[17] This is illustrated in Figure 2.

That the maximum points of U_N and p_K occur at a larger city size under assumption B should not be surprising. Under assumption B no capital rentals are spent on housing produced in cities and hence the amount of housing relative to X_1 produced is less under B than under A. Since under assumption B, the amount of homesites relative to X_1 produced is less, the rising resource costs of city size only offset the benefits of agglomeration as reflected in U_N and p_K at a larger city size.

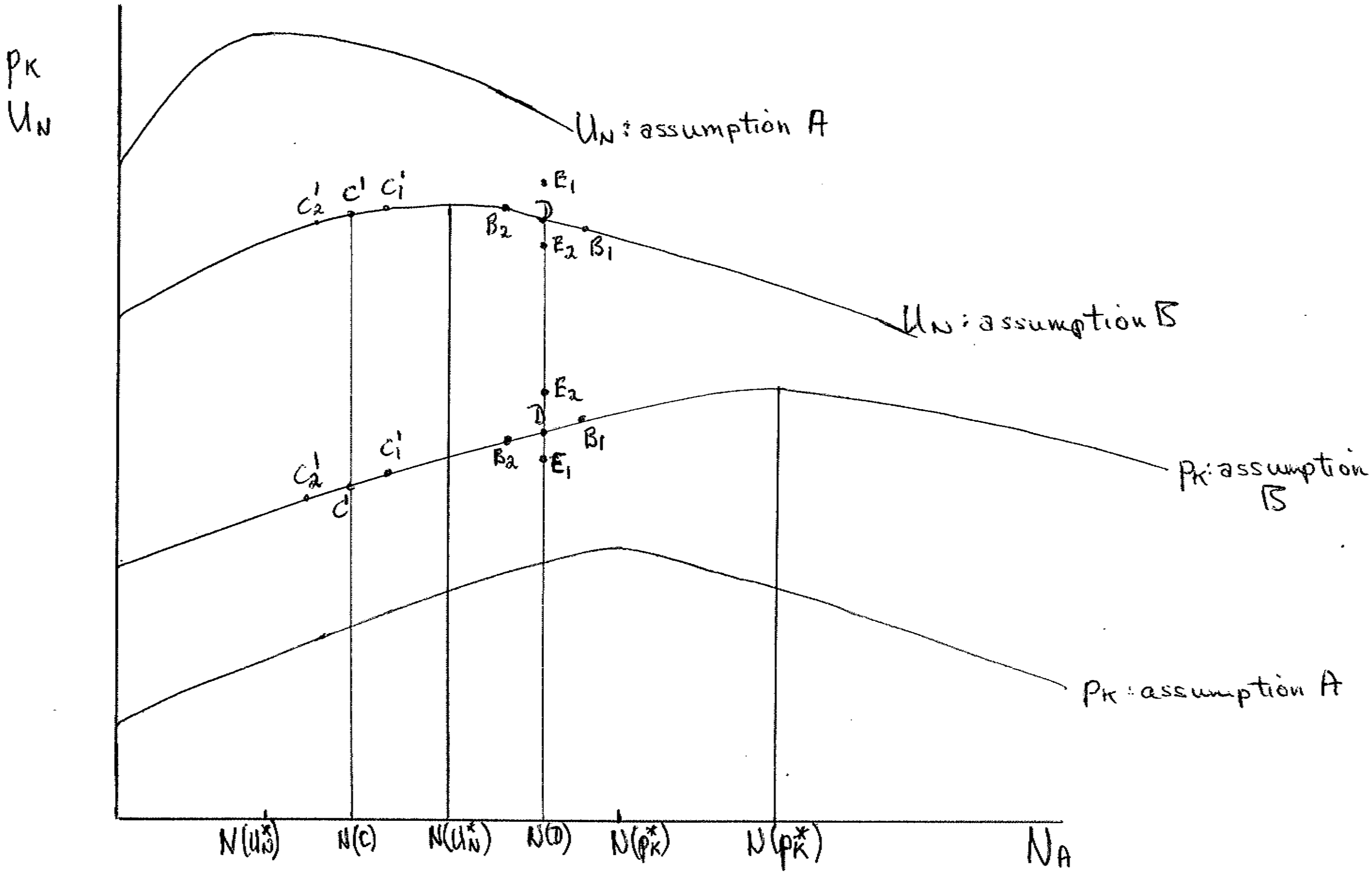

Figure 2: Utility and Capital Rental Paths under Assumptions B and A.

[17] From footnotes 9 and 10, we know $s < 1$ and $t > 1$. Hence $-\log s > \log t$.

2. CITY SIZE

In this section, the utility and capital rental paths derived in the previous section are used to solve for city size. Optimum city size is solved for by maximizing an appropriate sum of real factor rewards such that the potential consumption good bundle or welfare of consumers is indirectly maximized. Therefore it is appropriate to maximize either, under assumption A where capital owners live in cities, a sum of U_N and U_K or, under assumption B, a "sum" of U_N and p_K. By doing this, the welfare of the country's inhabitants is maximized. Equilibrium city size is solved for given factor and output market equilibrium where factors move to maximize the perceived returns from their services, U_N and p_K. It will be shown that factor and output markets do not necessarily generate sufficient information and market signals for factor movements to generate optimum size cities.

To initiate the process of city formation, we start with one city in the economy producing X_1 and then increase the size of the economy. This does not mean we have a growth model per se, since no savings behaviour, population growth function, or technological change are specified. It is an indirect or artificial method of solution that is necessitated by the nature of the problem. This is particularly true for the market equilibrium solutions and presents problems that will be discussed in evaluating such solutions. However the method of solution does yield correct optimal solutions and does serve to reveal the problems in the workings of market forces and signals when determining equilibrium city size.

In this section, it is assumed there is only one type of city in the economy, cities producing and exporting X_1; until the next section the analytical complexity of factors shifting between different types of cities in response to factor price changes is avoided. Given the shapes of factor reward paths, under either assumption A or B with respect to where capital owners live, as the initial city size increases, a second, then a third, and so on, city will form if both utility and capital rental paths have a finite maximum.

Stability Conditions: The Lower Round on City Size

Before the analysis of city size commences, stability conditions in factor markets must be examined. For the rest of the paper only stable city sizes will be considered. Stability considerations arise as soon as there are two or more cities. Factor market equilibrium will prevail only of factor rewards are equalized which occurs when all cities producing X_1 are of the same size. Stability prevails if a random movement of capital and labour from one city to another city generates market forces or factor movements returning the cities to their original sizes.

In Figure 2, we have two cities of size $N_A(C)$ under assumption . These correspond to a U_N level of C^1 and a p_K level of C^1 where U_N and p_K are the factor measures that labours move to maximize and capital owners move to maximize. Note that C^1 is on a rising part of both U_N and p_K paths. Suppose a random move of factors occurs such that a small amount of capital and labour (holding K_A/N_A constant) move from one city to the other. In the receiving city p_K and U_N rise to C_1^1; in the losing city they fall to C_2^1. This induces further factor flows into the initially receiving city. Hence two or more cities of size $N_A(C)$ is unstable.

Only if we are on a declining part of the U_N path and on either the rising or declining part of the p_K path will city size be stable with two or more cities.[18] Thus the first stable city size with two or more cities occurs at N_A (U_N^*).

[18] To show that two cities of size such that we are on a declining

Optimum City Size

Optimum City Size occurs when an appropriate sum of factor rewards is maximized. At any given city size, a change in city size may benefit both groups of factors, capital owners and labourers, or it may benefit one group of factors while making the other worse off. The change in city size is optimal if the gaining group of factors can compensate the other for its losses (if any). The optimum is a Pareto-optimum.

There is intially one city in the economy and this city size and economy is growing. We want to know when it is optimal to form a second, then a third, etc., city. For the initial discussion the K_A/N_A ratio and q_1 are held constant. In solving for optimum city size the assumptions A and B where capital owners live play a crucial role. First optimum city size under Assumption A is discussed.

Assumption A. Under the luxury of our current assumptions--capital ownership is evenly divided amongst labourers, capital is mobile independent of capital owners, and there is only one type of city in the economy--we can define a clear optimum city size under assumption A. As these assumptions are relaxed, the concept of an optimum city size under assumption A will become somewhat more hazy.

part of the U_N path but rising part of p_K path is stable, the following argument is employed. In Figure 2 we are at $N_A(D)$. A random movement of capital and labour occurs moving us to B_1 and B_2 respectively on the two paths. In the receiving city at B_2 p_K rises and U_N falls. This initiates further movements of capital into and labour out of the initially receiving city. Suppose we go back to $N_A(D)$ where p_K has risen to E_1 in the city with the lesser amount of capital and has fallen to E_2 in the other city. (This is due to the K_A/N_A ratio effect where $\partial p_K/\partial(K_A/N_A) < 0$ from equation (13), if $\rho_1 < \alpha_1 + \zeta_1$ which is true if both paths have a finite maximum since then $\rho_1 \leq \alpha_1$.) These p_K values, induce capital to leave the city with more of it and we will return to points D or the initial equilibrium. It should be noted that at E_1 and E_2 there were divergences between the U_N's in the two cities not discussed. A full argument of stability would need an adjustment mechanism. Our argument is sufficient if capital adjusts faster to price differentials than labour does.

Given they work as labourers, all capital owners live in cities and buy housing, paying the price of housing in their city. Although capital owners invest to maximize capital rentals, in solving for optimum city size, the consumption benefits they receive from spending their capital receipts should be maximized. Similarly the consumption benefits of wages to labourers should be maximized. Thus in solving for optimum city size, U_N and U_K are utilized. Note as previously mentioned (p. 14 and footnote 8), our assumptions specified above also allow us this luxury of dichotomizing labour income, welfare, and actions by income source.

If the U_N and U_K paths never achieve a finite maximum and rise indefinitely, a possibility that arose in section 1, there are no limits to the benefits of increasing city size. If the paths do achieve a finite maximum and then decline, eventually it benefits both groups of factor owners if a second city forms. Note from equations (14) and (15) that the U_N and U_K paths achieve a finite maximum at the same city size. We will see this means there is no conflict under assumption A between capital owners and labourers as to when it is optimal to form additional cities and what optimum city size is.

In Figure 3, holding K_A/N_A constant, as city size grows beyond $N(U_N^*, U_K^*)$, the city size of maximum U_N and U_K, a second city should form when N_A equals twice $N(U_N^*, U_K^*)$. To ensure factor price equalization between cities, we would then have two cities of size $N(U_N^*, U_K^*)$. If two cities formed before $2N(U_N^*, U_K^*)$, resulting in city sizes less than $N(U_N^*, U_K^*)$ on the rising part of the factor payment paths, stability would not prevail in factor markets. As the two cities of size $N(U_N^*, U_K^*)$ continue to grow a third city of size $N(U_N^*, U_K^*)$ should form from the two cities when they reach size $3/2\ N(U_N^*, U_K^*)$. In general a n+1 city should form when the n cities reach size $\frac{n+1}{n}$ N(U_N^*, U_K^*). If $n \to \infty$, which will be called the large sample case, city size will approach N(U_N^*, U_K^*) where U_N and U_K are maximized. From equations (14) and (15), N_A equals N(U_N^*, U_K^*) can be solved from

$$\log t - 1/m + \frac{(1-c)(\alpha_1-\rho_1) + c\alpha_3(1-\rho_1)}{m(\alpha_1-c\alpha_1-c\alpha_3(\rho_1-1))} N_A^{-m} - \log N_A = 0 \qquad (16)$$

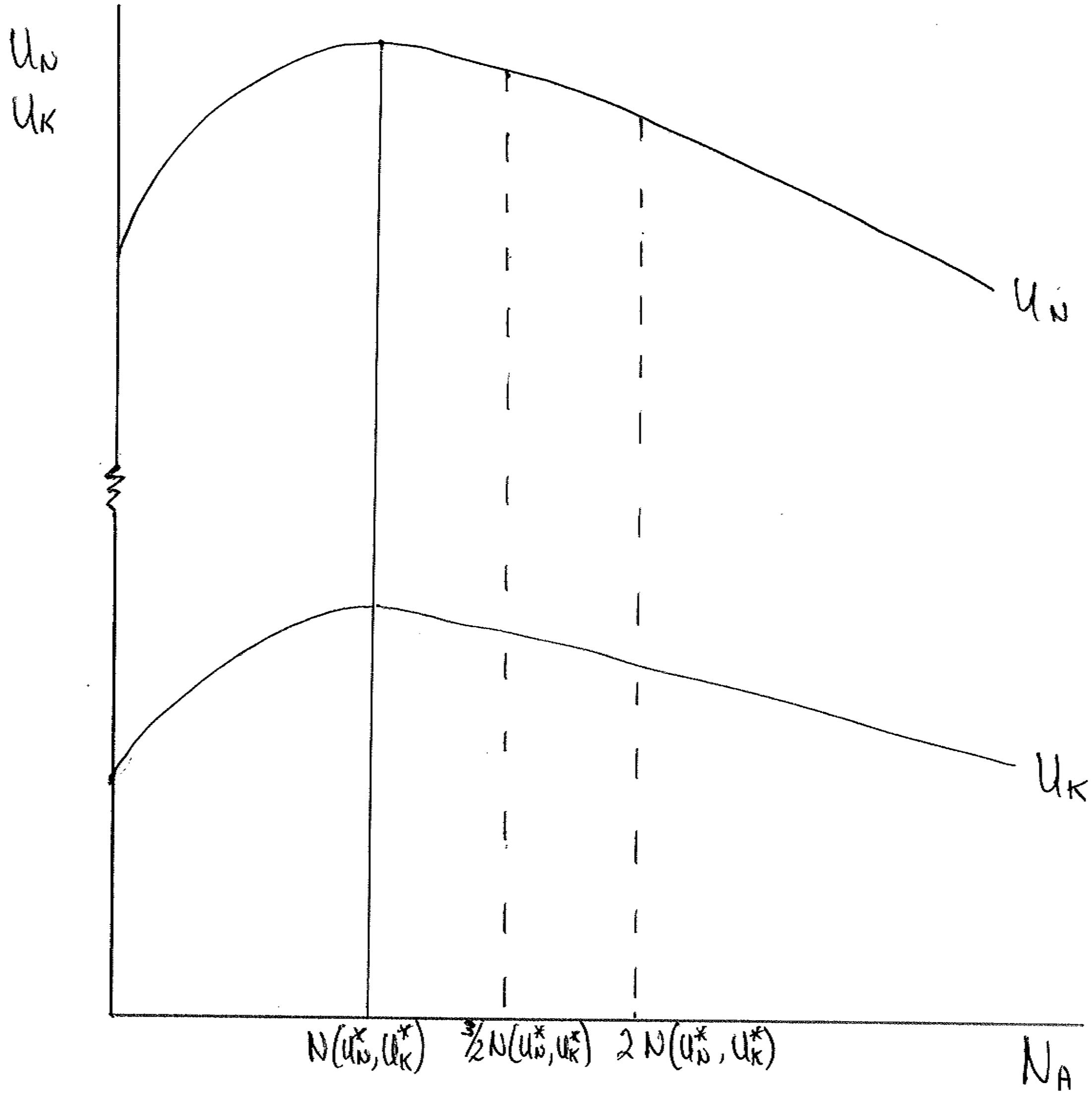

Figure 3: Optimum City Size: Assumption A

Note that a change in the K_A/N_A ratio as the economy grows would not affect city size. Regardless of K_A/N_A, U_N and U_K always attain a maximum at the same city size and hence optimal city sizes as well as equation (16) would be unaffected.

The city size $N(U_N^*, U_K^*)$ gives an indication of the maximum benefits of scale economies for our economy. The benefits of scale economies are limited because city size is limited by the existence of homesite production and decreasing returns to scale. Starting with one city in Figure 3, initially utility increases and then decreases. With two cities, utility falls again as the economy grows and declines until a third city forms. As the number of cities increases, the range of city sizes falls and the lower boundary of utility rises. For our large sample case we approach city size $N(U_N^*, U_K^*)$ and maximum welfare in the economy, as well as maximum benefits of scale economies. In a certain sense at $N(U_N^*, U_K^*)$, we approach a constant returns to scale case in production. Doubling the size of the economy would bring no further scale economy benefits.

Assumption B. Suppose now that capital owners live in the countryside or abroad in other countries. The price they pay for housing is independent of their investment returns and city size in this economy. Therefore their consumption benefits are maximized, when their capital rental receipts are maximized. To solve for optimum city size, the p_K and U_N paths are utilized.

If neither the p_K nor U_N paths have a finite maximum, factor rewards rise indefinitely as the initial city size increases and it never is beneficial to form a second city. If both paths achieve a finite maximum, then it will be beneficial to form multiple cities as the economy grows. The problem is it will be beneficial to form additional cities at different points for capital owners and labourers since from the previous section we know the U_N path achieves a maximum before the p_K path. For example, in Figure 4 as the number of cities in the economy grows very large, optimal city sizes for labourers and capital owners approach $N(U_N^*)$ and $N(p_K^*)$ respectively. How are these differences reconciled?

If the initial city size has reached twice $N(U_N^*)$ in Figure 4 labourers would be better off and capital owners worse off if two cities of size $N(U_N^*)$ formed. The size of the initial city should increase beyond twice $N(U_N^*)$ if capital owners can compensate labourers for their loss in utility of a second city not forming.[19]

Suppose the initial city size moves beyond twice $N(U_N^*)$ to $N(E)$ where it is optimal to a second city, yielding two cities of size $N(E^1)$. At $N(E)$, capital owners can no longer compensate labourers for not forming two cities of size $N(E^1)$. As illustrated in Figure 4, the loss to capital owners of two cities forming is $K_A(p_K(E)-p_K(E^1))$ and the gain to labourers is $N_A(U_N(E^1)-U_N(E))$. The compensation that could be offered by capital owners to individual labourers for not forming a second city is $M(K)$ where

$$M(K) = K_A/N_A(p_K(E)-p_K(E^1)) \tag{17}$$

The compensation demanded by a labourer for not forming a second city is $M(N)$ where

$$U_N(E^1) = a^a b^b c^c q_1^{-a} q_2^{-b} q_3^{-c}(p_N + M(N)) \tag{18}$$

p_N is the wage rate in city size E. Note that the calculation of p_N, q_3 and $U_N(E)$ would be affected by M since the demand for housing would rise as city income rose by $M(N)$. At $N(E)$ two cities of size $N(E^1)$ form because $M(N) \geq M(K)$.

[19] From the Coase [3] theorem, the same solution will be achieved if the compensation goes the other way--if labourers bribe capital owners.

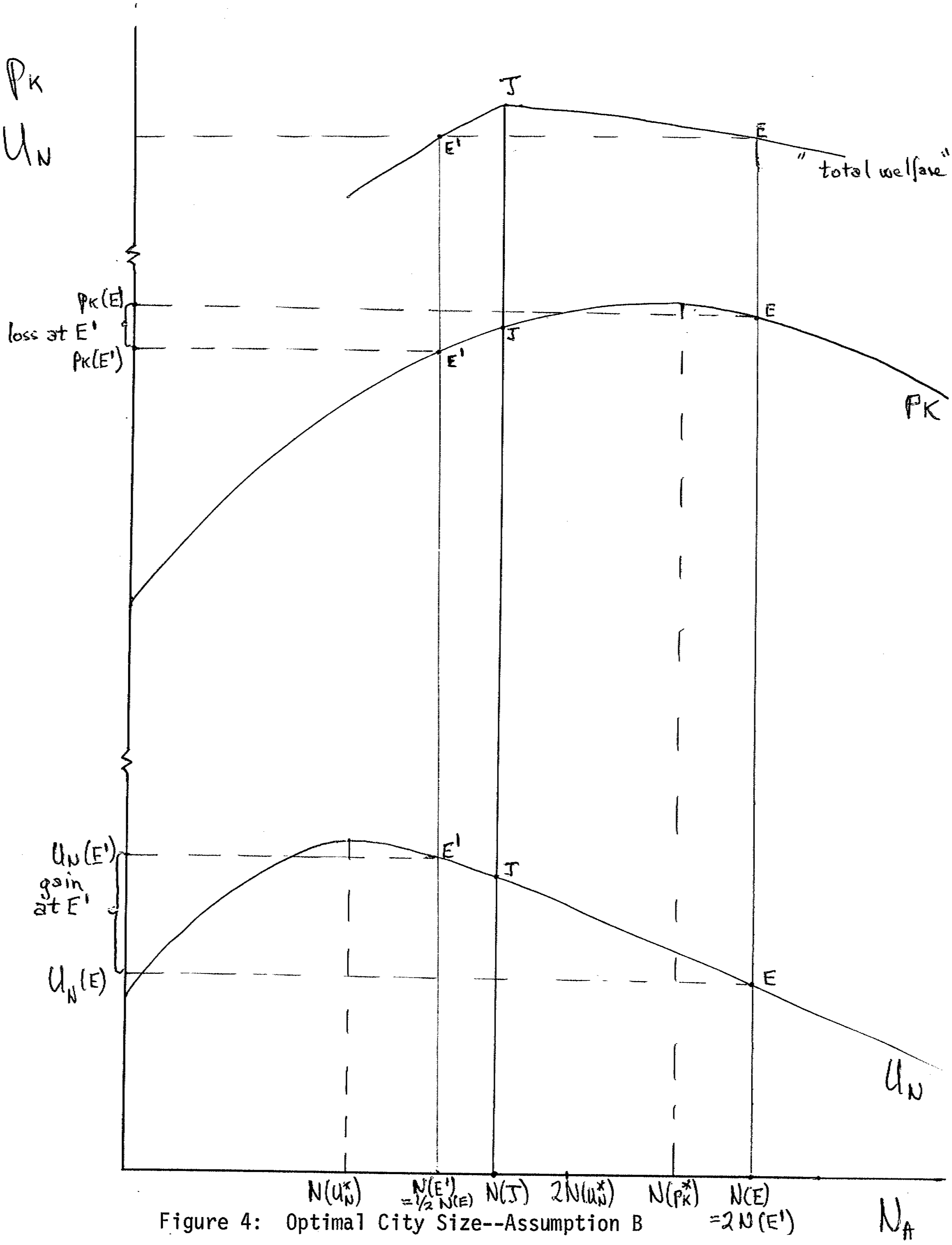

Figure 4: Optimal City Size--Assumption B

After two cities of size $N(E^1)$ form, the economy continues to grow with additional optimal size cities forming via our compensation mechanism. Of particular interest is the large sample case where the number of cities is very large and hence when an additional city forms the changes in city size of existing cities are minimal.

In the large sample case, an additional city forms when capital owners can no longer compensate labourers for their losses from not forming an additional city. Since city size changes are minimal when the additional city forms, the changes in U_N and p_K can be expressed in derivative form. In Figure 4 it is assumed at N(J) it is optimal to form an additional city and optimal city size approaches N(J). Note that $N(U_N^*) < N(J) < N(p_K^*)$. At J, $M(N) \overline{>} M(K)$ or from equations (17) and (18)

$$\left| a^a b^b c^c q_1^{-a} q_2^{-b} q_3^{-c} (dU/dN_A) \right| \overline{>} \left| K_A/N_A (dp_K/dN_A) \right| \qquad (19)$$

If we substitute in expressions for dU_K/dN_A and dp_K/dN_A from equations (13) and (15) this expression becomes:

$$\left| [C_N^{\prime} \frac{m(\alpha_1-\alpha_1 c-c\alpha_3(\rho_1-1))}{\rho_1-1}] \cdot [-\log s - 1/m + \frac{(1-c)(\alpha_1-\rho_1) + c\alpha_3(\rho_1-1)}{m(\alpha_1-c\alpha_1-c\alpha_3(\rho_1-1))} N_A^{-m} - \log N_A] \right|$$

$$\overline{>} \left| C_K^1 \frac{\alpha_1 m}{\rho_1-1} \; [-\log s - 1/m + \frac{\alpha_1-\rho_1}{\alpha_1 m} N_A^{-m} - \log N_A] \right| \qquad (20)$$

Solving (20), would yield $N_A = N(J)$, the optimum city size. Note that a change in the K_A/N_A ratio would not affect optimum city size or equation (20) just as it would not affect $N(U_N^*)$ and $N(p_K^*)$.

It has been argued that this process of compensation leads to an optimum city size in the sense that no changes in city size could occur that would make both groups of factors better off, allowing for compensation.[20] By this process of compensation we can effectively add the utility and capital rental paths weighted by N_A and K_A respectively to obtain an expression for total "welfare". In fact, by comparing equations (17) and (18) we are effectively adding them to obtain their maximum sum. The total welfare curve is represented in Figure 4. N(J) is our optimum city size for the large sample case. Also we show N(E) the point where it was optimal to form two cities of size $N(E^1)$ from one city.

Note that optimum city size under assumption B is larger than under A. First from section 1, we know that $N(U_N^*)$, the maximum point of the utility path, occurs at a larger city size under assumption B, because no capital rentals are spent on housing and homesites in the cities. Second, optimal city size occurs beyond $N(U_N^*)$ under assumption B, but at $N(U_N^*)$ under A (in the large sample case).

[20] As yet the traditional problem that arises when scale economies are external to the firm has not been discussed. This problem was raised previously on pages 5 and 6. To assert a true social optimum the problem must be dealt with. Throughout the rest of the paper, since this problem is independent of the analysis in the paper, it can be assumed either the problem has been taken care of or the optimum referred to is not quite the social optimum.

In section 1, when specifying production functions for traded good and homesite production, it was argued that social and private marginal costs would diverge because scale economies were externalities imposed upon the firm. The traditional solution to the problem is to offer subsidies and taxes to firms to eradicate the divergence between social and private marginal costs.

To do this, the production of L, the good with negative externalities, must be taxed at a rate of zp_L and the production of X_1 the good with positive externalities must be subsidized at the rate $\rho_1 q_1$. With respect to optimal city size, it is asserted without proof in this paper that the higher homesite costs shift the factor reward paths to the left whereas the higher return to traded good production shift the paths to the right. (The principle effect of the taxes is to change the values of all constants including t and s^{-1} in equations (13)-(15). The magnitude and direction of these changes is uncertain, see Henderson [6].) It is not clear which is the pervasive effect and whether optimal city size will be larger or smaller than the city size before taxation.

City Formation and Size: A Market Economy

We now solve for equilibrium city size in a market economy. In the initial analysis, we will see that market forces generate cities of sizes very different from the optimal sizes. In subsequent analysis under assumption B, we will be able to generate optimal size cities by introducing the concept of the city corporation, a theoretical institution parallel to land developers in the real world. Under assumption A with respect to where capital owners live, no theoretical institution or real world parallel can be devised to yield optimum city size in the market economy.

The market economy is characterized by atomistic behaviour of capital owners, firms, and labourers. Labourers when choosing a city to live in act to maximize their utility levels; this describes their market behaviour. Capital owners act to maximize investment returns under either assimption A or B with respect to where capital owners live. Even under assumption A, although their utility is affected by the cost of living in the city they live in, it is not affected generally by the cost of living in the city they invest in. In the national capital markets, they seek to maximize capital rentals. Thus in the solution of equilibrium city size and factor market equilibrium, the U_N and p_K paths are utilized under either assumption A or B.

However, it is the behaviour of entrepreneurs or firms that determines city size and city formation in our initial analysis. Starting with one city in the economy and increasing the size of the city, a second city will only form when an entrepreneur or firm sees it is profitable to leave the initial city and set up a second city. A second city is profitable when an entrepreneur can leave the initial city and initiate X_1 production in the countryside, paying competitive capital rents and labour utility levels.

The crucial point in analyzing firm behaviour is that, in the production of X_1 because scale economies are external to the firm, the individual firm acts unaware of any potential industry scale economies when making decisions. When moving to the countryside, the scale of operation is at the firm level and initially no industry scale economies will be experienced. From section 1, given individual firms have linear homogeneous production functions, firm size in our specification of X_1 production is indeterminant but is small enough to ensure a competitive industry. For simplicity, it is assumed that firms are of minimal size. The entrepreneur when moving to the countryside initially hires a unit of capital and labour in the new city dividing their services between X_1, X_3 and L production.

If both capital rental and utility paths rise indefinitely, an entrepreneur will never be able to competitively hire away capital and labour from the initial city, since factor rewards will always rise with city size. There will be only one city in the economy as in the discussion of optimal size.

If both utility and capital rental paths have a finite maximum, at some point, an entrepreneur will be able to profitably hire factors away from the initial city into a new city. In Figure 5, the firm can hire small amounts of capital and labour away from the initial city when it reaches size N(E). In the new city of size one, the entrepreneur will initially operate with a lower K_A/N_A ratio, explaining the shifts in the U_N and p_K paths relative to the paths for the larger city, as depicted in the Figure 5. The upward shift in the p_K path for the tiny city means that p_K in the city of one will be equal to or greater than p_K in the city of size N(E). The fall in the U_N path indicates that utility level in the small city is equal to or greater than level in the larger city where utility is a function of relative wage rates and city cost of living or housing. The equal to or greater than specifications are to allow for profits or return to entrepreneurship for formation of the new city.

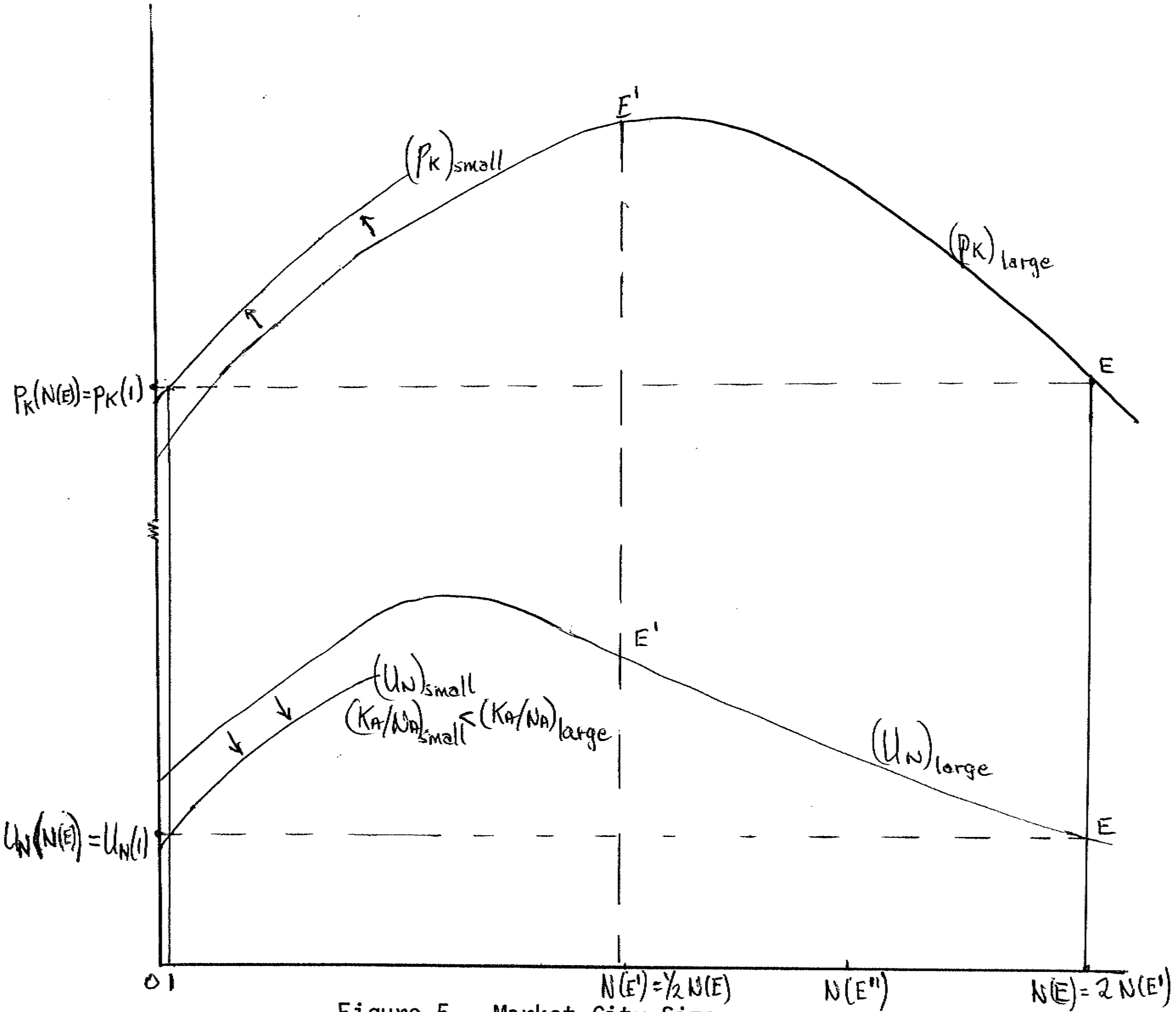

Figure 5. Market City Size

Now we have a city of size N(E) and one of size one[21]. Other firms

[21] Note that there is a speed of adjustment problem here. Suppose at E in Figure5 , a firm does not instantaneously go out and form a second miniature city. If our initial city size proceeds slightly beyond E then two or more separately located small firm/cities become profitable at a point beyond E. This raises the possibility of three cities forming from the initial one. To avoid this problem, we assume that a firm acts as soon as the initial city reaches size E.

will join the original entrepreneur in the new city induced by his current profits. Scale of output will increase and the city will move up its utility and capital rental paths inducing further factor flows. In final equilibrium there will be two equal size cities of size N(E') (where N(E) = 2N(E')) having the same K_A/N_A ratio. Note that capital rentals and utility levels are both higher at N(E') than at N(E).

At N(E') the two cities continue to grow until they both reach size N(E). At N(E), by the above process, a third city forms. The resulting equilibrium has three cities at N(E") where N(E") = 2/3 N(E). As the economy grows new cities continue to form with the lower bound on equilibrium city size approaching N(E), the point of city formation. Equilibrium city size is entirely different from optimum city size. For example, in the large sample case where the number of cities formally approaches infinity, equilibrium city size is at N(E) in Figure 5. Under assumption A, optimum city size is at $N(U_N^*)$ where $N(U_N^*)$ equals $N(U_K^*)$. Under assumption B, optimum city size lies between $N(U_N^*)$ and $N(p_K^*)$.

Is this divergence between equilibrium and optimum city size likely to persist in more sophisticated models and analysis? First technological improvements should occur as the economy grows, shifting the U_N and p_K paths out and up. If, for example, in Figure 5 we are at city size N(E) with a large number of cities, a technological change could occur so the paths shift (not shown in the figure) such that the new maximum point of the utility path ($N(U_N^*)$) occurs at city size N(E'). Given this interaction of technological change, it is not certain at a given point in time where on the paths we are at current city size.

Although this alleviates the impact, it does not solve the problem of inadequate market signals, arising from the fact that industry or even city scale economies are external to the firm and thus firms only act in the market when they would benefit from acting as a group, as in the optimum city size solution. These new cities formed by firms moving en masse from an old city to a new city, experiencing from the start significant industry or city scale economies.

However the divergences between optimum and equilibrium city sizes implies divergences between optimum and equilibrium capital rental and utility levels. These divergences could give rise to other market forces, if we add another participant or "actor" to our model, in addition to firms, labourers, and capital owners.

The City Corporation. Suppose we are at N(E) in Figure 6 in the large sample case. Optimum city size under assumption B is at N(J) as defined by equation (20). A move from N(E) to N(J) in city size would raise factor prices. In fact if by chance a city corporation were to form and to hire factors into a city of size restricted to be less than N(E), factor rewards that could be paid would rise in that city.

Suppose that a city corporation forms a city of size restricted to N(B). It pays slightly higher p_K and U_N than in the cities of size N(E) and guarantees higher earnings to firms. Doing this, it reaps profits per unit of capital or labour employed, approximately equal to the money equivalents of the bracketed amounts at N(B) in Figure 6. Other entrepreneurs will observe this profit and set up their own city corporations with cities restricted to size N(B). As the number of city corporations grows, they will compete for factors bidding up factor prices. The process only ends when factors are paid the value of their marginal products and there are no profits in the city corporation "industry". The city corporation "industry" is competitive.

But at N(B), profits could be made by further restricting city size.

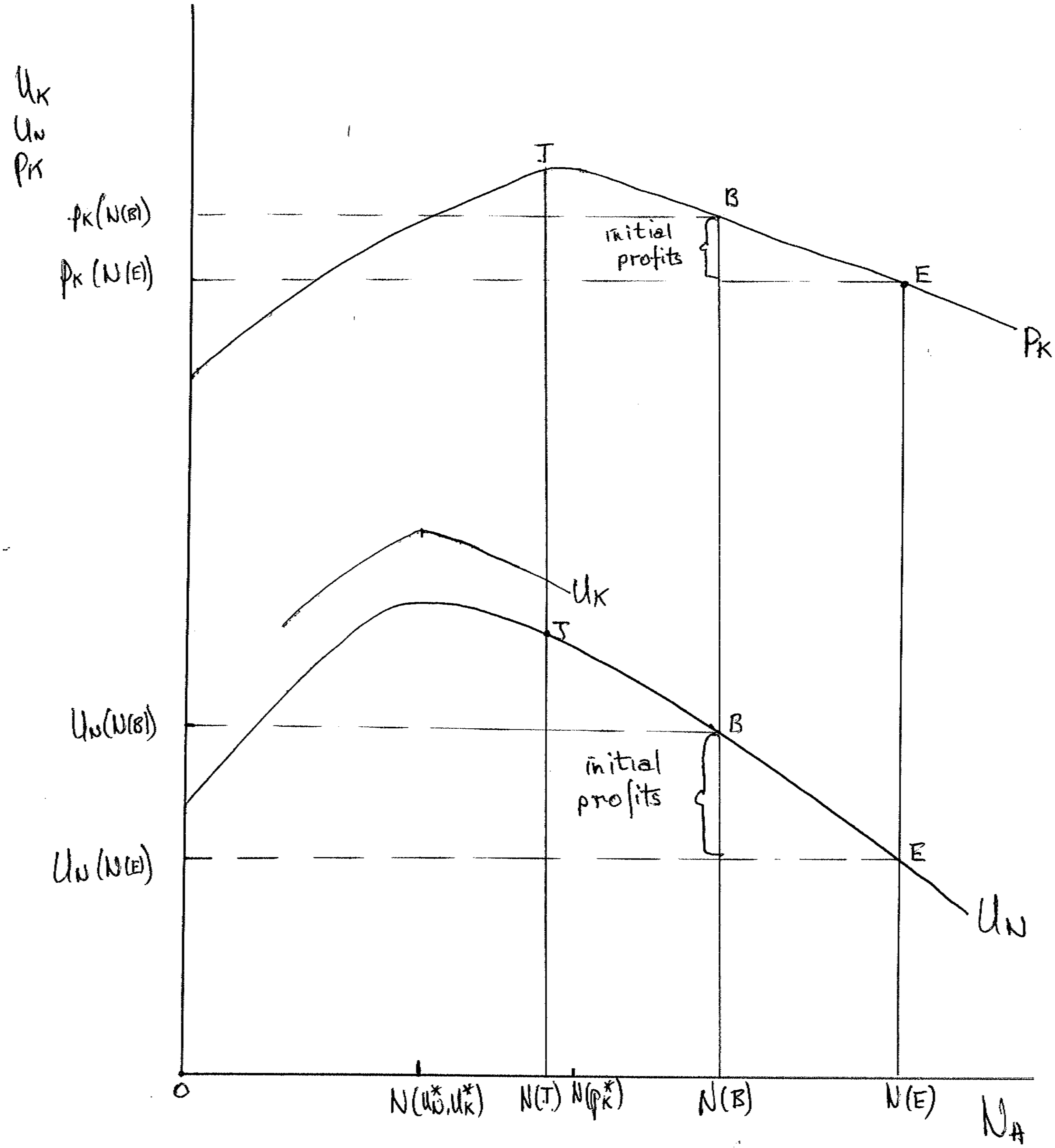

Figure 6. City Size--The City Corporation

Only at N(J), the optimum city size under assumption B could no more profits be made by restricting city size. As we saw in Figure 4 and equation (20), the sum of total factor rewards is maximized at N(J); and potential profits from changing city size and raising the value of marginal products above current factor payments no longer exist. The city corporation works "as if" the compensation mechanism used in the discussion of optimum city size is in effect. If our city corporation industry is competitive and has adequate information, we will approach city size N(J), the optimum city size under assumption B, but not under A.

Our system with capital owners, labourers, and firms had insufficient "actors" or market signals to achieve an efficient city size, so we added an additional "actor" or participant in the role of city corporations. This actor must be able to assess explicitly or implicitly the existence of scale economies and related cost of living in large cities. As the economy grows, he must act to form additional cities, en bloc, as demanded by our analysis of optimum city size. What could be the real world equivalents of the city corporation? One is city governments who would recognize the existence of industry scale economies and, having more knowledge than firms, act to effect with location taxes and subsidies a more optimal solution than the market solution.

A more realistic possibility is land developers. They could form new cities of restricted size and make an initial profit as the city corporation industry did. The profits would attract more land developers until an optimal number of cities is attained and no more profits are made. It is necessary though that the land developers initially control or own all the land in the city in order to be able to effectively restrict a city to the size they want.

It seems likely that land developers play a crucial role in the real world. In terms of our model they are essential to the achievement of optimum city size. In a more sophisticated model they would play a more intricate role. For example, if our model allowed for suburbs, land developers would form suburbs as our core type or Mills (1967) type city grew in size. Suburbs would allow for a) the release of pressure to form a completely new city due to rising commuting costs and b) a mechanism for a completely new (economic) city to form where the "suburb" or our new city would be economically independent of the old city (see p. 5 above). This type of city formation would be in addition to the type in our model, without suburbs.

Assumption A. It was stated above that, given city corporations, market forces would lead to a city size of N(J). This statement held regardless of whether assumption A or B pertained with respect to where capital owners lived. Capital owners when investing always seek to maximize capital rentals and therefore market city size is solved using p_K and U(N) paths yielding a city size of N(J). However under assumption A optimal city size is at $N(U_N^*, U_K^*)$ which is less than N(J).[22]

At N(J), there are no market forces leading us to the optimum city size at $N(U_N^*, U_K^*)$ in Figure 6. No more profits can be made by changing city size. By definition of N(J), the value of the gain in total utility to labourers from restricting city size below N(J) would be less than the loss in capital rentals to capital owners. Capital owners could not see that restricting city size throughout the economy would raise U_K in all cities,

[22] N(J) under assumption A is quantitatively smaller than under assumption B since $N(U_N^*)$and $N(p_K^*)$ for assumption A lie to the left of those points under assumption B (see Figure 3). N(J) under B is solved from equation (20); under A it is solved from the same equation except the constants C_N^1, C_K^1 and s^{-1} are replaced by C_N, C_K, and t.

although it reduced p_K. For example, if a city corporation formed one city of size $N(U_N^*, U_K^*)$, it would raise the utility level of capital owners and labourers living in the city. However the capital rentals the city corporation could pay out would simultaneously fall. All investors could earn higher capital rentals, which they are seeking to maximize, in cities of size N(J) and hence would not invest in a city of size $N(U_N^*, U_K^*)$.

The basic problem is investors act to maximize p_K, per se, not utility from spending p_K. There is no mechanism to ensure they will account for the cost of living given their location decision, when deciding what rate of return they should receive on their capital, regardless of where they invest. Is this an example of market failure? There is a "better" hypothetical solution than the market solution, but no way to obtain the solution, given the constraints of the functioning of markets.[23] In addition, as we relax our assumptions below the concept of an optimal city size under assumption A will become opaque.

3. EQUILIBRIUM IN THE ECONOMY

In this brief section a second type of city is introduced into the economy. This type of city produces homesites, housing, and a traded good X_2 produced with economies of scale. The type of city in the previous section

[23] If we constrained capital to move only when capital owners moved, then we would achieve our optimal solution because capital owners would account for the cost of living in making their dual investment-location decision, (see footnote 7 above). This achievement is illusionary and disappears if we relax our assumptions. First in the real world capital is more mobile than labour and if capital mobility was tied to labour mobility this would be devastating to economic growth, etc. Second, under our assumption of equal ownership of capital, when we introduce another type of city below, tying capital and labour mobility together would mean our two types of cities would have the same K/N ratio as the country. This again is a devastating restriction! Note the problem disappears if capital ownership and labour services are separated under assumption B.

specializing in X_1 production is a type A city; cities specializing in X_2 production are type B cities. A third or fourth type of city could be introduced each specializing in the production of a different traded good; but to develop the basic concepts only two types of cities are needed.

To have two types of cities in the economy it must be more efficient to have specialization than to have X_1 and X_2 produced in the same cities. Scale economies are more fully exploited by separating the production of X_1 and X_2 into different types of cities. Homesite production inefficiency rises by the same amount per unit of city labour input whether we separate the industries or have just one type of city. However, with specialization, scale economies are more fully exploited per unit of labour input since the labour input is not split in two ways but is concentrated in increasing the scale of one industry per city, rather than two. Note that for the case where there are multiple cities of each type, this implies that specialization results in a larger city size than non-specialization, because the point where the resource costs of further homesite production equals the resource benefits of further traded good production is at a larger output of traded goods.

Although it may seem paradoxical that larger city size and hence larger wastage in the production of homesites under specialization would be more efficient it is logical. Although there are higher prices and resource costs of homesites, utility and capital rentals are also higher due to the greater exploitation of scale economies in traded good production possible with specialization.

Given there are two types of cities, the conditions describing equilibrium in the economy will be examined. Also the process by which new cities form within a given type or group of cities must be re-examined, accounting for factor flows between the two types of cities.

Formal Derivation of Equilibrium in the Economy

First equilibrium in type B cities must be examined. Production conditions and equations in type B cities are the same as in type A cities except X_2 instead of X_1 is now produced. Consumption conditions are the same. Expressions for utility and capital rentals and the expressions for the paths of these variables are identical except all parameters subscripted 1 for X_1 coefficients are subscripted 2. Whether the factor reward paths attain a maximum and where they obtain the maximum is determined by the same equations, replacing X_1 production coefficients with X_2 coefficients. Given the shape of factor reward paths, the analysis of city size is the same, ignoring interactions between the two types of cities.

The following are the formal conditions for market equilibrium under either assumption A or B. In capital markets

$$(p_K)_A = (p_K)_B \tag{21}$$

where $(p_K)_A$ refers to capital rentals in type A cities. In labour markets

$$(U_N)_A = (U_N)_B \tag{22}$$

From equations (10) and (12), we know these variables can be expressed in terms of N_A, N_B, K_A, K_B, q_1 and q_2.

(Under assumption A, the labour market condition is actually $C[(q_3^{-c})_A(p_N)_A + \overline{p}_K\ (K/N)_A)] = C[(q_3^{-c})_B((p_N)_B + \overline{p}_K(K/N)_B)]$ where from equation (9), $C = a^a b^b c^c q_1^{-a} q_2^{-b}$. A labourer receives both labour and capital income. $\overline{p}_K$ is determined by his investment decisions and is exogenous to his location and consumption decisions. With only one type of city, q_3 is equal between all cities, so when U_N and p_K are equalized between cities so is $(p_N + \overline{p_K}\ K/N)q_3^{-c}$. With more than one type of city, q_3 varies between cities of different types when they are different sizes, so that $(p_N+\overline{p_K}\ K/N)q_3^{-c}$ will not be equal between cities when U_N (or $p_N q_3^{-c}$) is. In the larger cities p_N would have to be relatively greater in equation (22), so U_N is greater in the larger cities. Then $(p_N + \overline{p_K}\ K/N)q_3^{-c}$ will be equal between cities when $\overline{p_K}$ is. To induce a labourer to move to a high cost city, one must compensate him for his loss in value of deflated capital earnings. Note this distorts K/N allocation since all adjustment occurs through p_N and none through p_K (see below). This variation between types of cities is ignored here.)

Factor market equilibrium conditions are completed by:

$$n_a N_A + n_b N_B = N = \text{labour force in the country} \qquad (23)$$

$$n_a K_A + n_b K_B = K = \text{capital stock in the country} \qquad (24)$$

where n_a and n_b are the number of type A and B cities.[24]

Equilibrium in output markets is determined either by

$$q_1 = \overline{q}_1, \qquad q_2 = \overline{q}_2 \qquad (25)$$

in an open economy where trade prices are determined internationally, or by

$$q_2 = \overline{q}_2,\ \ q_1 = aY/X_1 n_a\ ,\ \ \text{where } Y = p_K(n_a K_A + n_b K_B) \qquad (26)$$

in a closed economy where q_2 is the numeraire and q_1 is determined from national demand conditions (see equation 8). (Note from footnote 6, X_1 in equation (26) can be expressed in terms of N_A and K_A; the actual equation is in Henderson [6].)

There are now six equations, (21) to (25) or (26), and eight unknowns,

[24] If capital owners live in the countryside under assumption B, some allocation of capital and labour would have to be made to provide countryside housing for capital owners. If they live in other countries, we are assuming that the allocation of capital to our economy is fixed. This very rigid assumption could be relaxed to having capital supplied according to some supply elasticity. Doing this would complicate the analysis with little gain in understanding the main issues presented in the paper. Formally, the model would need an equation such as $K = f(p_K)$.

K_A, K_B, N_A, N_B, q_1, q_2, n_a, and n_b. Two more equations, one each describing the process of city formation within types A and B cities, complete the information needed to solve for equilibrium. These equations are not derived in this paper but are discussed in the sub-section below.

To solve the system of equations describing equilibrium would require an iterative rather than simultaneous solution due to the algebraic complexity of the two equations describing the number of type A and B cities. For example, suppose equation (25) is in effect. Then, for each allocation of capital and labour to type A and type B cities, such that economy resources are exhausted and equations (23) and (24) are satisfied, we have an equilibrium number of type A and B cities and city sizes as determined by our equations for city formation discussed below. For one of these factor allocations between type A and B cities and the resulting city sizes, we also satisfy the factor market equations, (21) and (22). Thus we achieve equilibrium.

City Size Determination Revisited

Here city size formation is briefly re-examined given there are two or more types of cities in the economy; a formal presentation of the subject is in Henderson [6]. For expositional ease, our discussion assumes the existence of the city corporation mechanism and atomistic behaviour of firms. This implies the economy functions "as if" the compensation mechanism used to solve for optimum city size in section 2 is in effect. This further implies that, under assumption B where capital owners live outside the cities, the market economy achieves optimum city size. Under assumption A, for reasons discussed above, the market economy does not achieve optimum city size.

(Note that our concept of optimum city size under assumption A has become somewhat more hazy. Labourers in their role as capital owners invest to equalize p_K between cities. They locate to equalize $(q_3^{-c}(p_N + \overline{p_K}\ K/N)$

where $\overline{p_K}$ is fixed in capital markets but p_N varies between cities (see pp. 41-42 above). If they invested and moved to maximize and equalize $q_3^{-c}(p_N+p_K\ K/N)$ where both p_N and p_K would vary according to consumer location, they would be better off. The principle is still that capital owners should invest to maximize utility from spending capital receipts not capital receipts per se. Since the cost of living varies between cities labourers owning capital should demand different rates of return on capital invested in *a* location depending on *where they live*. This would have the parallel effect in Figure 6 of limiting city size to maximize capital rentals deflated by the cost of living in the economy cities plus utility from labour earnings. That is, people would account for their cost of living when making investment decisions. Under the market system, the effect is to maximize rentals not utility, as well as to distort city K/N ratios to equalize $q^{-c}(p_N + \overline{p_K}\ K/N)$ between cities (see p. 42 above). Achieving the "optimal" system is only possible in a hypothetical world ignoring market constraints. In the real world it would imply an impossible phenomenon--two people investing in city b would be paid differentially because they lived in cities c and d with different costs of living. Note the problem disappears under assumption B. Capital owners, not constrained by labour location decisions, pick the least expensive place to live such as the countryside, the Bahamas, or Majorca!)

Suppose we start with one each of type A and B cities in the economy, both in equilibrium on declining parts of their utility paths. Capital rentals and utility levels are equalized between cities.[25] Equilibrium output and

[25] If both cities are on the rising part of their utility paths, equilibrium may be unstable. Once the first city is beyond its maximum point on the utility path, then two types of cities can exist with stability. See Henderson [6] on this point.

factor rewards in each city are a function of relative traded good prices (q_1 and q_2), relative factor endowments (K/N), relative degrees of increasing returns to scale (ρ_1/ρ_2), and factor intensities in production (α_1, δ_1, β_1 etc.). As the economy continues to grow, at some point total welfare (the "sum" of $U_N N + p_K K$ - see pp. 27-31 above) for the economy is maximized by a second type A or B city forming. The derivation of this optimal point is the same as previously except now if, say, a second type A city forms not only will resources be diverted from the first to the second type A city but resources will be diverted from the type B city to type A cities. Equations can be derived for type A and type B cities showing when it is optimal, given total economy resources, to form new cities within each type (see Henderson [6]).

As demonstrated in section 2, for the large sample case where the number of type A and B cities is very large, when an additional city forms the flow of resources to the new city from other <u>individual</u> type A and B cities is minimal. An additional type A city forms when capital owners can no longer compensate labourers for not forming an additional type A city (capital rentals are rising and utility levels falling with increases in city sizes). Therefore parallel to equation (19) for the large sample case is equation (27) where an additional city should form when

$$\left| a^a b^b c^c q_1^{-a}\ q_2^{-b}\ q_3^{-c}\ (dU/dN_A) \right| = \left| K/N\ (dp_K/dN_A) \right| \qquad (27)$$

Note that (27) is weighted by K/N, the economy factor endowment ratio, not K_A/N_A tye type A city factor allocation ratio.[26] Equation (27) and the

[26] Equation (27) can be reduced parallel to equation (20) to yield: $K_A/N_A \left| C_N \frac{m(-\alpha_1+c\alpha_1+c\alpha_3(\rho_1-1)}{\rho_1-1} [\log t - 1/m + \frac{(1-c)(\alpha_1-\rho_1) + c\alpha_3(1-\rho_1)}{m(-\alpha_1+c\alpha_1+c\alpha_3(\rho_1-1))} - \log N_A] \right| =$ $K/N \left| C_K \frac{\alpha_1 m}{\rho_1-1} [\log t - 1/m + \frac{\alpha_1+\rho_1}{\alpha_1 m} N_A^{-m} + \log N_A] \right|$. Note that city size is a

parallel equation for type B cities complete the missing equations in the above sub-section needed to solve for economy equilibrium.

4. EXTENSIONS OF THE MODEL

The model can easily be extended to incorporate the effect of natural resources, agriculture, and transportation costs in trade. Unfortunately, space limitations make this the topic of a later paper. In brief, natural resources, agriculture, and transportation costs provide for the existence of new types of cities such as processing and extraction centres, ports, break-in-bulk points, etc., all specializing in certain activities. They also give rise to a geographic pattern of cities and regions and place the model in a more "realistic" light. All the principles of city formation, size, and type derived above remain intact. In addition, city size is further limited beyond the reasons due to rising commuting and congestion costs by the extent of the market for its traded good or transportation costs of the good. A city may now split into two cities before would be predicted from the commuting cost arguments alone, to effect transport cost reductions in trade because the new city would locate in another region of the country lowering that region's costs of obtaining the export good of that type of city.

function of both K/N (the weights attached to (economy and city) factor reward changes with a change in type A city size) and K_A/N_A (the factor ratio effect on U_N/p_K in type A cities).

REFERENCES

[1] Alonso, W. Location and Land Use. Cambridge: Harvard University Press, 1964.

[2] Beckmann, M. Location Theory. New York: Random House, 1968.

[3] Coase, R.N. "The Problem of Social Cost", Journal of Law and Economics, Vol. 3, 1961.

[4] Borts, G.H., and J.L. Stein. Economic Growth in Free Market. New York: Columbia University Press, 1964.

[5] Chipman, J.S. "External Economies of Scale and Competitive Equilibrium", Quarterly Journal of Economics, Vol. 84, No. 3, 1970, pp. 347-385.

[6] Henderson, J.V. "The Types and Sizes of Cities: A General Equilibrium Model", unpublished Ph.D. dissertation, University of Chicago, 1972.

[7] Herberg, H., and M.C. Kemp. "Some Implications of Variable Returns to Scale", Canadian Journal of Economics, Vol. 2, No. 3, 1968, pp. 403-15.

[8] Jones, R.M. "The Structure of Simple General Equilibrium Models", Journal of Political Economy, Vol. 73, No. 6, 1965, pp. 261-272.

[9] Isard, W. Location and Space Economy. Cambridge: Massachusetts Institute of Technology Press, 1960.

[10] ________, et al. General Theory: Social, Political, Economic, and Regional. Cambridge: Massachusetts Institute of Technology Press, 1969.

[11] Lösch, A. The Economics of Location. New Haven: Yale University Press, 1954.

[12] Melvin, J.R. "Increasing Returns to Scale as a Determinant of Trade", Canadian Journal of Economics, Vol. 2, No. 3, 1968, pp. 124-31.

[13] Mills, E.S. "An Aggregative Model of Resource Allocation in a Metropolitan Area", American Economic Review, Vol. 57, No. 2, 1967, pp. 197-210.

[14] ________. "The Value of Urban Land", The Quality of the Urban Environment, edited by H.S. Perloff, Resources for the Future, Inc., Baltimore: The Johns Hopkins Press, 1969, pp. 231-53.

[15] Mundell, R.A. International Economics, New York: The MacMillan Co., 1968.

[16] Muth, R. Cities and Housing. Chicago: University of Chicago Press, 1969.

[17] Pearce, I.F. International Trade: Book Two. London: MacMillan and Co., Ltd., 1970.

[18] Tiebout, C. "The Pure Theory of Local Expenditures", Journal of Political Economy, Vol. 64, 1966, pp. 416-424.

[19] Vernon, R. "The Produce Cycle", Quarterly Journal of Economics, 1965, pp. 190-207.

[20] Weber, A. On the Location of Industry. Chicago: University of Chicago Press, 1929.