TIMELINESS, TRADE AND AGGLOMERATION

James Harrigan
Anthony J. Venables

Working Paper 10404
http://www.nber.org/papers/w10404

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2004

Produced as part of the Globalization programme of the UK ESRC funded Centre for Economic Performance at the LSE. Thanks to Niko Matouschek and participants in seminars at the LSE, the IMF, University of Nottingham, City University of Hong Kong, and AEA meetings (San Diego) for helpful comments. The views expressed in this paper are those of the authors, and do not necessarily reflect the position of the Federal Reserve Bank of New York, the Board of Governors of the Federal Reserve System, or the National Bureau of Economic Research.

Timeliness, Trade and Agglomeration
James Harrigan and Anthony J. Venables
NBER Working Paper No. 10404
March 2004
JEL No. F1, L0, R12

ABSTRACT

An important element of the cost of distance is time taken in delivering final and intermediate goods. We argue that time costs are qualitatively different from direct monetary costs such as freight charges. The difference arises because of uncertainty. Unsynchronised deliveries can disrupt production, and delivery time can force producers to order components before demand and cost uncertainties are resolved. Using several related models we show that this generates hitherto unexplored incentives for clustering. If final assembly takes place in two locations and component production has increasing returns to scale, then component production will tend to cluster around just one of the assembly plants.

James Harrigan
International Research Department
Federal Reserve Bank of New York
33 Liberty Street
New York, NY 10045
and NBER
james.harrigan@ny.frb.org

Anthony J. Venables
Department of Economics
London School of Economics
Houghton Street
London WC2A 2AE
a.j.venables@lse.ac.uk

1. Introduction

People pay a lot of money to save time. A modern economy is inconceivable without air travel and air shipment, which are ways of saving time at the expense of money. For workers in urban areas the main component of commuting costs is time. For international trade in manufactured goods estimates of the costs of the time-in-transit range as high as 0.5% of the value of goods shipped, *per day* (Hummels 2001). Protagonists of 'just-in-time' manufacturing techniques emphasize the importance of organizing and locating production to ensure timely delivery of parts and components.

Surprisingly, these observations have had little impact on the economic analysis of location decisions[1]. Economists have worked with an aggregate of 'transport costs' or 'trade costs' to capture the penalty of distance, while simply remarking that these costs are a shorthand for a wide range of penalties (e.g. Fujita, Krugman, and Venables 1999). These include freight and other monetary transactions costs; lack of information about markets and suppliers and about local institutions and regulations; difficulty in monitoring contracts; the impossibility of face-to-face contact and communication; and the fact that distance introduces delay into completion of trades. It is unlikely that summarizing these penalties as a single value of 'trade costs' is adequate for understanding their effects. The objective of this paper is to contribute to the process of unpacking and evaluating the different elements of these costs[2].

We focus on the costs associated with delivery times and argue that timeliness is not only a quantitatively important aspect of proximity, but also

[1] For instance, Fujita and Thisse's (2002) lucid and authoritative new book doesn't have "time" in the index. There is some modeling of the issues in Harrigan and Evans (2002) and Venables (2001).

[2] Previous attempts to evaluate non-monetary trade costs include study of the benefits of face-to-face contact, see Leamer and Storper (2001) and Storper and Venables (2003).

matters qualitatively, creating an incentive for clustering of activities. The context is the time taken between initiating a project and completing it and making delivery to the consumer. We suppose that physical distance between stages of the production and selling process (e.g. between component manufacture and final assembly, or between production and final sale) slows down that process, and argue that slowing down matters for several reasons. One reason is discounting and other analogous factors, such as the physical depreciation or technical obsolescence that component parts may be subject to during shipment. These costs will not be the focus of our attention, although we note that they may be large – computer chips become obsolescent very rapidly, so it is not sensible to ship them on the slow boat.

Other reasons why delay matters are intimately connected with uncertainty. One set of arguments has to do with the synchronization of activities; production cannot be completed until all the parts have arrived, so uncertain arrival times of components can have a cost that is quite disproportionate to the cost of any single component. Other arguments arise since uncertainty – about demand or costs – makes it profitable to postpone production until as much uncertainty as possible has been resolved. The cost of delivery delay is that it brings forward the date at which production decisions have to be made, orders placed, and expenditures incurred. We look at the implications of this when there is uncertainty about the product characteristics that are demanded, and when there is uncertainty about the total level of demand or costs.

Of course, saving time is always going to be beneficial, just as is saving freight charges. To make the point that there may be qualitative (as well as a quantitative) implications of timeliness, we develop all our models in a very particular framework that enables us to assess the profitability of clustering activities together. The framework is one in which there are two locations, each of which has an assembly plant or retailer supplying local final demand. The assembly process uses a number of component parts, and increasing returns in production of these components are sufficiently great that each is produced in a single plant. Where do the component producers locate? Clustered around one of the assembly plants, or divided between the two locations? We show that the demand for timeliness in delivery creates a force for clustering of plants around a single assembler/retailer. This is a previously unexplored agglomeration mechanism.

We develop this argument in a series of models. Section 3 outlines a benchmark case in which there are monetary trade costs, but delivery is instantaneous and component producers do not cluster. In section 4 we look at the issues raised by the synchronization of delivery of components, and show that uncertain delivery times will cause clustering of component producers. Sections 5 and 6 show how uncertainty about demand and or about production costs are also forces for clustering. Section 5 also investigates the role of inventories. However, before developing these models we briefly connect our to approach to the extensive management literature on just-in-time (JIT) production.

2. Just-in-time

The just-in-time (JIT) approach was pioneered by Toyota Motors in the 1950s. Its main features are that components are delivered in small but frequent batches, that minimal stocks are held, and that ‘quantity control is built in’. The perceived advantages are a reduction in the cost of holding stock, rapid response to customer orders, and the ability to rapidly detect and fix or replace defective components. Effective implementation of JIT is thought to require long term supplier/ customer relationships and, where possible, proximity. In the management literature on JIT production it has been suggested that the spread of

JIT systems might be expected to lead to a geographical reconcentration of supplier firms and customers (e.g. Dicken 1998).

The importance of proximity is illustrated by the example of General Electric's appliances division in their attempt to implement JIT in the 1980s and 90s. They were hampered by the fact that some suppliers were several thousand kilometres away from GE plants, this causing a 1993 decision to increase inventory levels (Jones, George and Hill 2000). The US auto-industry has been extensively studied, although identifying the effects of JIT on supplier location is a tricky empirical question. Assemblers tend to locate where suppliers are already located, and in addition there are non-JIT reasons why suppliers may want to be near assemblers (such as minimizing transport costs irrespective of timeliness considerations). Klier (1999) assembles a comprehensive dataset on assemblers and suppliers and shows that, since the advent of JIT, new supplier plants are more likely to locate near their assembly plant customers than they were before the advent of JIT. Klier also finds that proximity generally means "within a days drive", rather than right next door, which implies that the agglomeration force of JIT operates at the regional rather than the urban level. This is consistent with the results of Rosenthal and Strange (2001), who find "...shipping-oriented attributes (manufactured inputs, resources, perishability) influencing agglomeration at the state level..." (pg. 193).

Our goal in this paper is to develop some simple models that capture some of the features referred to in this literature, and to draw out their implications for the spatial concentration of activity.

3. A timeless model

We develop our ideas in a family of models, each based on two *a priori* identical locations *A* and *B*, where final assembly occurs and demand is met. Assemblers in *A* and *B* require components, and component suppliers can be located in either *A* or

B. We refer to the final producers as assemblers but the idea is more general: "assemblers" could be service firms who require a variety of manufactured or service inputs, or retailers who sell a variety of products.

Assemblers produce with constant returns to scale and the final assembled product is non-tradeable, so assembly must take place in both locations. Components are tradeable, although trade typically takes time. The number of types of components, *N*, is fixed, and production of each incurs a plant level fixed cost and then has constant marginal cost. The fixed cost is large enough to ensure that each component is only produced in one location, either *A* or *B*, and our primary question is to ask where this component production takes place.

In most of the models we develop all components are necessary to production of the final product, raising the question of how surplus is split between independent assemblers and component producers. The theory of (non-cooperative) bargaining offers no answer to this when there is more than one supplier (see Sutton 1986, Binmore and Dasgupta 1987). However, it does lead us to expect that the outcome will be efficient, maximizing the combined returns to all parties. In our analysis we therefore look merely at the total returns to different locational patterns, and not at how these might be divided between players. Our focus on maximizing the total surplus could also be rationalized by assuming that all a single firm controls all activities, or that both the assembly plants are controlled by a single firm, and all components produced by a single firm. Nash bargaining by these two parties over the location of components' production and division of the surplus (by setting prices for components) would then lead to the efficient outcome.

Before we move to our models of timeliness, it is instructive to look briefly at a timeless case based on ingredients from a standard economic geography model. In this benchmark model, assemblers in *A* and *B* each combine *N* symmetric inputs in a CES production function to create a unit of final output. The value of producing one unit of final output in location *A* is revenue minus the costs of producing and shipping components,

$$V_A = p - \left[N_A r_A^{1-\sigma} + N_B \left(\tau r_B \right)^{1-\sigma} \right]^{\frac{1}{1-\sigma}} \tag{1}$$

The exogenously given price of final output, net of any assembly costs, is *p*. The remaining term is the cost of the parts required to produce a unit of output. The cost function has elasticity of substitution σ; N_A is the number of components sourced locally with unit production costs r_A, while the remaining N_B ($N_A + N_B = N$) come from the other location with unit production cost r_B and shipping cost factor $\tau > 1$. Notice that, since we are looking for efficient outcomes, we use the unit production costs of components, r_i, which may not be the same as the prices at which they are traded. Furthermore, we will henceforth refer to V_A as the profits of assembly in *A*, noting that it is both the profits of the assembler and profits (before fixed costs) earned by component producers on supply of parts to *A*. A similar equation gives profits in *B*.

What values of N_A and N_B maximize total profit, $V_A + V_B$? The total number of component suppliers is fixed at *N*, so that $N_A = N - N_B$, and we let input costs be the same in each location. Making these substitutions in (1) and taking the derivative of V_A with respect to N_B gives

$$\frac{\partial V_A}{\partial N_B} = \frac{r(\tau^{1-\sigma} - 1)}{\sigma - 1} \left[N + N_B(\tau^{1-\sigma} - 1) \right]^{\frac{\sigma}{1-\sigma}} < 0 \tag{2}$$

$$\frac{\partial^2 V_A}{\partial N_B^2} = \frac{-r\sigma(\tau^{1-\sigma} - 1)^2}{(1-\sigma)^2} \left[N + N_B(\tau^{1-\sigma} - 1) \right]^{\frac{2\sigma-1}{\sigma-1}} < 0 \tag{3}$$

These derivatives establish that V_A is decreasing and concave in N_B: shifting assembler locations from *A* to *B* has an increasingly negative effect on the returns to assembling in *A*. The opposite is true for the returns from assembly in *B*. The point here is that the *increasing* marginal cost of remoteness implies that the sum $V_A + V_B$ is maximized when half of the suppliers locate in each region. Equivalently, there is a decreasing marginal value to proximity. The point is illustrated in Figure 1, in which there are 10 components, and the number of component suppliers located in *B* is on the horizontal axis[3]. Curves give profits in each place, and their sum, $V_A + V_B$, is maximized when $N_A = N_B$. A lower elasticity of substitution, σ, gives less curvature and a flatter $V_A + V_B$ schedule, but only in the limit, when the elasticity of substitution is zero, do the curves become linear, and their sum horizontal.

This result does not turn on a CES cost function. Quite generally, if the assembler did not adjust its input quantities as N_A and N_B changed, then V_A would be the straight line joining values of V_A at $N_A = 0$ and $N_B = 0$. The possibility of adjustment means that V_A lies on or above this straight line, as illustrated. More formally, consider any symmetric unit cost function $c(\mathbf{r},\mathbf{r+dr})$ in which inputs are partitioned into a group (*A*) available at price *r,* the remainder available at price $r + dr$ (group *B*). Quantities demanded in each group are $x_A, x_B, x_A \geq x_B$. The increase in costs when a product moves from group *A* to group *B* is $\frac{\partial c}{\partial r}dr = x_A dr$ (by Shepherd's lemma). As more products enter group *B* so x_A must increase (in order that input levels are sufficient to produce the unit of output[4]), meaning that the cost of moving inputs from group *A* to group *B* is increasing. This increasing marginal costs gives the convexity of the cost function with respect to N_B, and the consequent concavity of profits.

[3] This and other figures are generated by simulation of the models. Parameter values are given in the appendix.

[4] The cross-partial derivatives of a symmetric unit cost function are positive, so raising the price of some inputs increases demand for other.

The conclusion is therefore that, in this benchmark case, there is no incentive for clustering of activity. The *ex ante* symmetric locations, *A* and *B*, are also symmetric *ex post*, as component producers are split in equal numbers between the locations. With this benchmark in mind we now turn to models where remote supply incurs a time cost.

4. Synchronization

Our first model of timeliness turns on uncertainty about delivery time, and the consequent risk that production may be delayed by the late arrival of components from a distant supplier. We model this by supposing that each assembly firm seeks to produce a unit of output for delivery at a particular date. Assembly uses labour to combine *N* different component parts into final output using a Leontief production function with unit coefficients. Of course, production cannot be completed until all the parts needed have arrived.[5] For the moment, we assume that holding stocks of components is infeasible or prohibitively costly. This might be because of very high storage or depreciation costs, or simply because the exact specification of the product is unknown prior to the decision to produce, an idea we pursue in the next section.

Transport of components between locations is costless, but timely delivery of parts can only be guaranteed if the assembler and parts supplier are located in the same region. The probability of timely delivery is $q < 1$ if supplier and assembler are located in different regions. Assuming that delivery of each part is iid across suppliers and assemblers, for assemblers located in *A*,

$$\Pr(\textit{all parts arrive on time}) = q^{N_B}$$

$$\Pr(\textit{at least one part arrives late}) = 1 - q^{N_B}$$

[5] This production function is formally identical to Kremer's (1993) o-ring technology.

where as before N_B is the number of parts suppliers located in B, $N_A + N_B = N$. Clearly, Pr(*all parts arrive on time*) is decreasing in N_B and (importantly, as it turns out) convex in N_B:

$$\frac{\partial q^{N_B}}{\partial N_B} = q^{N_B} \ln q < 0, \quad \frac{\partial^2 q^{N_B}}{\partial N_B^2} = q^{N_B} \left[\ln q\right]^2 > 0. \qquad (4)$$

This means that each part which changes from being supplied locally to remotely decreases the probability that all parts arrive on time, but does so at a diminishing rate. The intuition for this is straightforward: if one part is delayed, it doesn't matter if a second part is also delayed.

There are several reasons why delays in completing assembly might be bad for profits. One is demand decay. Many goods and services have demand which peaks at a certain time and the price that the assembler can get for the final product falls unless it is delivered on time. Another is that some assembly costs have to be met whether production occurs or not. For example, if labor must be hired to assemble parts, then wages must be paid regardless of whether all parts have arrived. Think of labour as a cost which must be incurred before the outcome of the delivery process is known, so that if there are delays, labour must be hired again once all parts arrive.

To capture these arguments, let final demand be characterized by a reservation price which is p on the day that demand is realized and $p(1-\delta)$ one day later, $\delta \in (0,1)$. Profits if all parts are delivered on time are therefore

$$v_A^0 = p - \beta w_A - N_A r_A - N_B r_B \qquad (5)$$

where β is the daily unit labour requirement for parts assembly and w_A is the wage. If parts are delivered one day late, the reservation price falls and labour must again be hired, so profits are

$$v_A^1 = p\left(1-\delta\right) - 2\beta w_A - N_A r_A - N_B r_B \qquad (6)$$

The difference between profits on day 0 and on day 1, $\delta p + \beta w_A$, is the penalty paid by firms who suffer late delivery of parts. Expected profits are just profits if there is no delay minus the expected cost of delay,

$$V_A = v_A^0 - \left(1 - q^{N_B}\right)\left(\delta p + \beta w_A\right) \qquad (7)$$

If there are no cost differences between the two locations, then (5) and (7) imply that expected profits in *A* are decreasing and convex in N_B: the hit to expected profits of sourcing an additional part from far away gets smaller as the number of them increases.

Symmetric results apply to expected profits in *B*, which has the important implication that total expected profits are maximized at $N_B = 0$ and at $N_B = N$. This is illustrated in Figure 2. In contrast to the benchmark model of the previous section, total expected profits are *minimized* at $N_B = N/2 = N_A$: with such a division of production, neither suppliers in *A* nor in *B* get the benefit of reliable deliveries. This illustrates the increasing marginal value of timeliness: if almost all parts have guaranteed on-time delivery, an increase in share of timely parts has a bigger effect on expected profits than if most parts are subject to erratic delays. As a result, there is an economic force leading to the agglomeration of all suppliers in either *A* or *B*.

The point of this simple case is then, that although the locations are ex ante symmetric, the efficient location of component producers is asymmetric. It is best to have one assembler operating in a cluster of all the component suppliers and producing without delay, while the other bears the full cost of the uncertainties associated with delivery delay. How does this compare to the benchmark model? In that model, as the elasticity of substitution σ goes to zero the cost function becomes linear, as it is in this model, and the $V_A + V_B$ line in Figure 1 becomes flat. The difference from Figure 2 arises because in the synchronization model of this section there is a fixed cost associated with delay so that (in expected value) there is a fixed cost associated with not sourcing from nearby suppliers. In the benchmark model, there is no such fixed cost to sourcing remotely.

Finally, notice that in the synchronization model the difference between locations shows up as a productivity difference. One of the key facts about agglomeration is that localized industries have higher measured productivity (see Rosenthal and Strange, 2003, for a review of the evidence). The model offers an explanation for this: localized activities benefit from timeliness, which reduces or eliminates periods when production is interrupted by delayed delivery. If all suppliers locate in *A*, then assemblers in *A* never have to pay labour twice, while assemblers in *B* have to pay labour a second time with probability 1 - q^N. Since output is the same in each location, relative productivity in *A* is given by the ratio of expected unit costs:

$$TFP_{AB} = \frac{(2 - q^N)\beta w + rN}{\beta w + rN} > 1 \qquad (8)$$

This TFP advantage for assemblers in *A* is increasing in the probability that at least one part is delayed and in the importance of assembly labour in total costs. It is also increasing in the total number of parts, which might be thought of as complexity[6].This is intuitive, since the greater the number of parts the greater the chance of a delay in having all parts arrive. This result suggests that parts used in more complex activities have a greater incentive to cluster than do parts used in simpler activities.

[6] To show this, we calculate the derivative of TFP with respect to *N*, holding labor's share in cost fixed (this requires an offsetting drop in *r* as *N* increases so that *rN* is constant, that is, *Ndr* + *rdN* = 0). The result is $\frac{\partial TFP_{AB}}{\partial N} = \frac{-\beta w}{\beta w + rN} q^N \log q > 0$.

5. Inventories and product specification uncertainty.

In the previous section's model, we ruled out holding parts' inventories as a solution to the assembler's problem. In this section we relax this, and address the tradeoff between just-in-time ordering and holding inventories. It turns out that inventories can be a substitute for timely delivery, but only if the cost of holding inventories is small enough relative to the benefit.

One aspect of demand uncertainty is uncertainty about precisely which final product is demanded. The assembler's problem is to quickly produce final output once the exact specifications of demand are observed, and he can do this in two ways. The first is to hold large inventories of parts, so that he can assemble whatever ends up being desired. The second is to have parts' suppliers right next door who can quickly produce and deliver the parts needed for the desired final product. By assumption, faraway suppliers can not deliver quickly enough to fill orders placed after demand is observed.

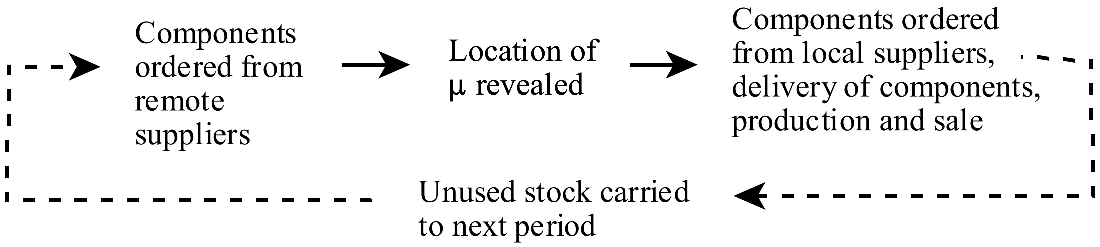
As in the previous section, the final good is assembled using *N* components, each of which has fixed coefficients in production. However, each component comes in a continuum of varieties (the unit interval), differentiated in the eyes of consumers. Thus, car engines come in different sizes, bodywork in different colours, etc. There is a unit mass of consumers, each consuming one unit of the good, and their preferences for varieties of each component may be either high or low, but are unpredictable. We model this unpredictability with the parameter $\mu \in [0,1]$, which we call "pickiness". For each component, consumers have high preference for a subset of measure μ of varieties and low preference for the remainder $1 - \mu$. Smaller values of μ correspond to more picky tastes, while larger values correspond to less discriminating tastes.

Consumer tastes are illustrated in Figure 3, in which the horizontal axis is the measure of varieties for one component. Preferred varieties lie in a set of measure μ; these varieties need not form a connected set although, for simplicity, this case is illustrated. The vertical axis is quantity, and the rectangle ABCD has height $1/\mu$. This is the number of units of each high preference variety demanded by the unit mass of consumers. Thus, the pickier are consumers the narrower is their set of preferred varieties μ and hence the greater the level of demand for each of these varieties, $1/\mu$. This pattern is repeated for each component, and we assume that there is no correlation between demand for the varieties of different components; preferences over engine size are uncorrelated with preferences over exterior colour.

The assembled final good is valued by consumers as a function of how closely each of its *N* component parts match the consumer's preferences. Products that are 'perfect' – all their components having a high preference variety – have reservation price $\overline{p}$. Those that have *k* low preference components (*k* 'mismatches') have reservation price $\overline{p}\phi(k)$, $\phi(0) = 1$. We assume that $\phi(k)$ is decreasing and convex in *k*, implying that the price falls at diminishing rate with multiple mismatches. This convexity assumption is satisfied if consumers' willingness to pay for the product as a whole is a standard CES function of the utility of individual components, as is demonstrated in the appendix. As a consequence of this convexity, a firm will produce as many perfect products as it can, concentrating all its mismatched components in as few a products as possible, rather than spreading its mismatched components over many units of final output.

The problem for assemblers is that consumers are both picky and fickle. Assemblers know the value of μ, but its location (in the unit interval of possible varieties) is initially unknown. For example, car assemblers know that μ exterior colours will be popular, but they do not know which ones. The more picky consumers are (the smaller is μ), the more difficult it is to cater to their whims, since the popular characteristics are harder to predict.

Delivery of components from remote suppliers takes time, so these components have to be ordered before the assembler knows the exact specification of demand – the location of μ. This is then revealed, components are ordered from local suppliers, and delivery and production take place. Any unused components can be carried over to the next period at a cost, and the production cycle is repeated indefinitely. This gives the following time line:


Components ordered from remote suppliers
Location of µ revealed
Components ordered from local suppliers, delivery of components, production and sale
Unused stock carried to next period

Given this timing, what quantities of what varieties of each component should the assembler in A be holding when production commences? For each locally supplied component the assembler knows the μ varieties that have high preference, and orders quantity $1/\mu$ of each of these. For each remotely supplied component, s denotes the quantity of each variety ordered in advance and held when production commences. If $s = 1$ then only μ consumers can be served with their preferred variety of the component, the remainder $(1 - \mu)$ having to make do with a low preference variety; all $s = 1$ units of each component get used up in production. If $s > 1$, then $s\mu$ consumers get their favorite variety and the remaining $(1 - s\mu)$ make do with low preference ones. Since one unit of each component gets used in production each period, stock of $s - 1$ is carried into the next production cycle, with new purchases replacing stock used.[7] Notice that it is never optimal to have $s < 1$, as one unit is required to produce the one unit of output, nor $s \geq 1/\mu$, as this is sufficient to perfectly match consumer preferences.

The per-period profit from assembly at A equals the revenue from selling perfect and imperfect products, minus the cost of carrying inventories over to the next period, minus the cost of purchased inputs (as before, we ignore the costs of assembly labour). This profit is

$$V_A = s\mu\overline{p} + (1 - s\mu)\overline{p}\phi(N_B) - (s-1)\gamma N_B - N_A r_A - N_B r_B \tag{9}$$

The first term says that, with stock level s, $s\mu$ consumers can receive products that perfectly match their preference. Other products contain mismatch in all N_B of their remotely supplied inputs, so are valued at $\overline{p}\phi(N_B)$. Remaining terms in the expression give the costs of producing the components and the inventory cost, where γ is the unit cost of holding stock to the next production cycle, so $(s - 1)\gamma N_B$ is the total cost of inventories.

Efficiency is achieved by choosing s to maximize (9). The problem is linear in s, so the assembler either chooses $s = 1/\mu$, to be able to perfectly match demand, or chooses $s = 1$, the minimum required to produce one unit of output regardless of specification. At $s = 1/\mu$, all the products sold by the assembler perfectly match consumer demands, at the cost of having unused components that are carried forward into the next period, with carrying charges on these inventories. While at $s = 1$ some of the products sold contain low preference components and hence sell for lower price, but the seller doesn't need to carry inventories forward. Evaluating V_A at these points, we find the profit-maximizing choices are

[7] For each of the preferred varieties, stock carried forward is zero. For each of the 1- μ low preference varieties the volume of stock carried forward is s - (1 - $s\mu$)/(1 - μ) = (s - 1)/(1 - μ).

$$\begin{aligned} s &= s_{\max} = 1/\mu && \text{if} \quad \overline{p}\mu\left[1-\phi\left(N_B\right)\right] > \gamma N_B \\ s &= s_{\text{minx}} = 1 && \text{otherwise} \end{aligned} \tag{10}$$

Outcomes are illustrated on figure 4, for the case with $r_A = r_B$. The horizontal axis is N_B ($N_A = N - N_B$), the intersecting dashed curves give profit when maximal and minimal levels of stock are held, and the maximized value is the upper envelope of these curves, max{$V_A(s{=}1/\mu)$, $V_A(s{=}1)$} with switch point as indicated in equation (10). To the left of the switch point maximal stock levels are held and V_A is linear in N_B. To the right of the switch point stock holding becomes too expensive an option, so firms set $s = 1$, producing a standard range of products that are not tailored to preferences; in this range V_A is convex. The upper envelope of these curves is convex, and hence their sum, $V_A + V_B$, is as illustrated. Efficiency is achieved by putting all suppliers in one place. One assembler has all components supplied locally and produces 'customized' products that perfectly match demand; it does this without holding stocks, because it relies on the proximity of suppliers. The assembler in the other location chooses not to hold stocks, instead preferring to produce a product range which is less well tailored to consumer demand.

Under what circumstances does the configuration illustrated in figure 4 apply? For some parameters it is profitable to hold stocks for all values of N_B, since $V_A(s{=}1/\mu) > V_A(s{=}1)$ over the entire range. In this case the sum $V_A + V_B$ is horizontal and the location of component producers is indeterminate. This case arises only if, evaluating the critical condition at $N_B = N$, $\overline{p}\mu\left[1-\phi\left(N\right)\right] > \gamma N$. If this condition is not satisfied then either there is an interior solution, as illustrated, or stockholding is never profitable, $V_A(s=1) > V_A(s=1/\mu)$ for all values of N_B. In both these case the sum $V_A + V_B$ is convex and clustering is the efficient outcome.

The factors that are conducive to clustering are now clear. Clustering occurs if $\frac{N}{\mu} > \frac{\overline{p}\left[1-\phi(N)\right]}{\gamma}$. The left hand side measures the quantity of stocks that ensure the firm can meet demand. It is larger the greater is N, the number of component types for which stock may have to be held, and the smaller is μ, i.e. the pickier are consumers. The right hand side gives the cost of not holding stocks relative to the cost of holding them, so a small value of this makes stockholding less desirable. It is smaller the higher is γ, the direct cost of holding stocks, and the smaller is the return when stocks are not held, $\overline{p}\left[1-\phi\left(N_B\right)\right]$. Convexity of the function $\phi(k)$ is also required for the $V_A(s=1)$ function to be convex and for there to be an intersection of the curves; as in the preceding section, we therefore require that the first mismatch be more expensive than the second, and so on. The appendix gives full details for the case of CES preferences.

The model of this section therefore gives two main messages. One is that, even if the direct costs of holding a unit of stock (γ) are not that high, the fact that stocks need to be held over a wide range of varieties of components can make the stock-holding strategy expensive. The other is that uncertainty about demand specification coupled with time in transit can generate clustering. Putting these together, we see that in industries where products are complex (a high N) or demand is unpredictable (low μ), the presence of time in transit will induce an equilibrium with clustering. One location will contain all the suppliers and produce customized products; the other has to import components, and produces a standard range of products. Although physical productivity is the same in both locations, the output price and hence the value of output per unit of input is higher in the location with the cluster. This will cause measured TFP to be higher in the cluster.

6. Demand (or cost) uncertainty.

In the preceding models incurring failures (late delivery or mismatched components) becomes progressively less costly, and it is this that gives the potential convexity of the profit functions and a motive for clustering. We now turn to an alternative model in which the mechanism is somewhat different. There is no uncertainty about the arrival time of components, or about the composition of demand. Instead, there is simply uncertainty about the level of demand, and the location of plants affects the extent to which it is possible to react to information about the position of the demand curve. In this way, we build on the work of Evans and Harrigan (2002), who examined a model of "lean retailing" and its implications for international specialization.[8] We develop the model with demand uncertainty, although show at the end of the section that assembly cost uncertainty has identical effects.

Demand for the output of each assembler can be high or low, represented by a linear inverse demand curve in which the intercept depends on the state of nature, so

$$p_i = \alpha^s - \beta y_i\,, \qquad i = A, B, \qquad s = H, L, \qquad \alpha^H > \alpha^L \tag{11}$$

where p_i is price and y_i is quantity of final product in region i, and superscript denotes the state of nature. High demand occurs with probability ρ. Whether high or low, demand is fleeting, and falls to zero if not met immediately.

[8] This section goes beyond their model in focusing on the location of multiple input suppliers.

As before, the production function has fixed unit input coefficients for each component, and we ignore labour costs in assembly. The assembler in region *A* faces the following sequence of decisions. First, she has to choose the quantity x_B of components to order from each of the N_B remote suppliers. These have to be ordered before the state of nature is revealed if they are to arrive in time for production. The state of nature is then revealed, and firms choose quantities of components x_A^s from each of the local suppliers. Finally, delivery of all components takes place and production occurs. This is summarized by the following time line:

Choose x_B: → α^s revealed: → Choose x_A^s: → Produce $y_A^s = \min\left[x_B, x_A^s\right]$.

The assembler's second stage choice problem (once the state of nature, $s = H, L$, is known) is to choose x_A^s to maximize v_A^s, defined as

$$v_A^s = x_A^s\left(\alpha^s - \beta x_A^s\right) - N_A r_A x_A^s, \qquad s.t.\ x_A^s \le x_B, \qquad s = H, L. \tag{12}$$

The maximand is revenue (where we have used the production function and the inverse demand curve) minus the costs of locally supplied inputs. The constraint reflects the fact that the assembler will never choose more local components than the quantity set by the supply of components coming from region *B*, because of the fixed coefficient technology. We solve this problem by maximizing the Lagrangean

$$L_A^s = x_A^s\left(\alpha^s - \beta x_A^s\right) - N_A r_A x_A^s + \lambda^s\left[x_B - x_A^s\right], \qquad s = H, L. \tag{13}$$

The first order condition with respect to x_A^s implies,

$$\lambda^s = \alpha^s - 2\beta x_A^s - N_A r_A, \qquad s = H, L. \tag{14}$$

There are two qualitatively different outcomes, depending on parameters including the level of demand. In one, production is constrained by the quantity of components coming from the remote supplier, so $x_A^s = x_B$ and $\lambda^s > 0$. In the other this constraint does not bind so $\lambda^s = 0$ and x_A^s is solved from (14); some components ordered from *B* are unused and freely disposed of.

The assembler's first stage problem is to choose x_B, before the state of nature is known, to maximize expected profits

$$V_A = \rho v_A^H + (1-\rho) v_A^L - N_B r_B x_B. \tag{15}$$

Varying x_B changes costs directly, and also changes v_A^H and v_A^L via the inequality constraint in (12). The first order condition for this problem is

$$\frac{\partial V_A}{\partial x_B} = \rho\lambda^H + (1-\rho)\lambda^L - N_B r_B = 0$$

since the Lagrange multiplier measures the value to the objective of a unit relaxation of the constraint.

As noted above, there are two cases to study. One we call the no-flexibility case, in which production in both states is constrained by the quantity of components supplied by remote producers. Output is therefore the same in both states, independent of the realization of demand, $x_A^L = x_A^H = x_B$, and $\lambda^H > 0$, $\lambda^L > 0$. The other is the flexibility case in which, if demand is high, production is constrained by the supply of pre-ordered components, so $x_A^H = x_B$ and $\lambda^H > 0$. However, if demand is low then not all these components are used, so $x_A^L < x_B$ and $\lambda^L = 0$. There is free disposal of unused components.[9] Which regime applies depends on parameters, including the values of N_A and N_B. We look first at the flexibility case, then turn to the no-flexibility case and the boundary between the regimes.

[9] Obviously, it is not profitable to discard components in both the high and the low state. The assumption of free disposal could be replaced by costly stock holding into a future period.

In the flexibility case, solution of first order conditions (14) and (16) gives,

$$x_A^H = x_B = \frac{1}{2\beta}\left(\alpha^H - N_A r_A - \frac{N_B r_B}{\rho}\right), \qquad \lambda^H = \frac{N_B r_B}{\rho},$$
$$x_A^L = \frac{1}{2\beta}\left(\alpha^L - N_A r_A\right) < x_B, \qquad \lambda^L = 0. \tag{17}$$

Using these equations we can show that the inequality $x_B > x_A^L$ holds providing $N_B r_B < \rho\left(\alpha^H - \alpha^L\right)$, this condition defining the boundary of the flexibility regime.

In this flexibility case, how do output levels and profits depend on the location of input producers? Increasing N_B has the effect of decreasing output in the high demand state and increasing it in the low state, which is to say that output is less responsive to demand the more components are produced far away. Formally, using $N_A = N - N_B$, setting $r_A = r_B = r$ and differentiating (17),

$$\frac{dy_A^H}{dN_B} = \frac{dx_B}{dN_B} = \frac{r(\rho - 1)}{2\rho\beta} < 0, \qquad \frac{dy_A^L}{dN_B} = \frac{dx_A^L}{dN_B} = \frac{r}{2\rho\beta} > 0. \tag{18}$$

The explanation is that higher N_B increases the total number of components left unutilized, and hence the expected cost of production; this reduces the profit maximizing level of output in the high demand state, y_A^H. However higher N_B also means that, if the low demand state transpires, a higher proportion of inputs have zero shadow price (the components from region B which, at the margin, are discarded); this reduces the marginal cost of production in the low state, so increasing quantity produced, y_A^L.

The effects of varying N_B on profits are given by differentiating (15). Using (12), (14) and (17) with $r_A = r_B = r$ gives (see appendix for derivation):

$$\frac{dV_A}{dN_B} = \rho\frac{dv_A^H}{dN_B} + (1-\rho)\frac{dv_A^L}{dN_B} - N_B r\frac{dx_B}{dN_B} - r x_B = r(1-\rho)\left[x_A^L - x_B\right] < 0. \tag{19}$$

The loss of profits due to a marginal increase in N_B is simply the expected cost of the quantity of this component that remains unused. Using the values of x_A^L and x_B given in (17), we further derive,

$$\frac{dV_A}{dN_B} = \frac{r(1-\rho)}{2\beta}\left[\alpha^L - \alpha^H + \frac{N_B r}{\rho}\right] < 0, \qquad \frac{d^2V_A}{dN_B^2} = \frac{r^2(1-\rho)}{2\rho\beta} > 0. \qquad (20)$$

Increasing N_B therefore reduces profits at a decreasing rate, by the convexity of V_A. The intuition is that if the country A assembler did not adjust its production plan then V_A would decline linearly, as more types of component are discarded in the low state. Adjustment raises profits, giving the convexity. As we have seen, increasing N_B decreases output in the high demand state and increases it in the low state, so reducing the gap between x_B and x_A^L. The implication is that when parameters are such that firms behave flexibly (that is, sell a different amount depending on the state of demand), there is a force for clustering of component suppliers around one of the final assemblers.

In the no-flexibility case, solution of first order conditions (14) and (16) gives

$$\begin{aligned} x_A^L &= x_A^H = x_B = \frac{1}{2\beta}\left[\rho\alpha^H + (1-\rho)\alpha^L - N_A r_A - N_B r_B\right] \\ \lambda^L &= N_B r_B - \rho(\alpha_H - \alpha_L) \qquad\qquad\qquad\qquad (21) \\ \lambda^H &= N_B r_B + (1-\rho)(\alpha_H - \alpha_L) \end{aligned}$$

The first equation gives purchases of components and hence also the level of output. This is the same in both states, so demand variability goes entirely into the price. Expected profits, V_A, can be computed using (21) in (12) and (15). For present purposes, the important point to notice is that if $r_A = r_B$ then output and sales levels do not depend on location of assemblers (the division of N between $N_A = N_B$, see equation (21)), so neither do profits. In the interior of this regime having more local component suppliers does not induce the assembler to change its behavior, and profits are constant.

The boundary between the flexible and non-flexible regimes is where the shadow value of x_B in the low state, λ^L, is zero, i.e. $N_B r_B = \rho(\alpha_H - \alpha_L)$. This is of course the same condition that gives the edge of the flexibility regime, where output levels in the high and low demand states just become equal, as discussed after equation (17).

The complete picture is illustrated in figure 5. The horizontal axis gives N_B, and the vertical axis gives levels of production and profits of the country A assembler. The no-flexibility regime is where $N_B r_B > \rho(\alpha_H - \alpha_L)$; a sufficiently large number of components come from remote suppliers that it is very costly to leave some of each of them unused when the low state occurs. By contrast, when $N_B r_B < \rho(\alpha_H - \alpha_L)$ then only a small share of component types face the risk of being left unutilized and discarded. It is therefore worthwhile to order a larger quantity of each type of remote component, x_B, output becomes state contingent, and the flexibility case applies.

Notice that there are now two distinct arguments creating convexity of profits, V_A, with respect to N_B. One is that, within the flexibility regime, profits are convex, as discussed above. The other arises because of the kink in V_A due to the change in regimes. Intuitively, having more local suppliers is of no value until some threshold is passed – only then is it worth adjusting production to exploit the benefits of rapid delivery times. The implication is, once again, that there is an incentive for all input suppliers to cluster in one location. In such a situation, one of the assemblers becomes completely flexible, ordering all its inputs from local suppliers once the level of demand is known. The other is inflexible, as all its inputs take time to be delivered and must be ordered before the state of nature is known.

Several other remarks are worth making on this model. First, price variability is lower in the location with the cluster of activity, as quantities are responding to demand shocks. With linear demands the expected price is the same in *A* and *B*, $Ep_i = \rho p_i^H + (1-\rho)\, p_i^L = \frac{\overline{\alpha} + Nr}{2}$, as is the expected quantity sold, $Ey_i = yp_i^H + (1-\rho)\, y_i^L = \frac{\overline{\alpha} - Nr}{2\beta}$. However, since the region with the cluster produces more in the higher price state, the average value of output produced *E(p*$_i$*y*$_i$*)/Ey*$_i$ is higher in the region with the cluster. Once again, the location with the cluster has higher measured productivity.

Finally, notice that this structure is isomorphic to a model in which shocks are on the cost side, rather than the demand side. Suppose that revenue $x_A^s\left(\alpha^s - \beta x_A^s\right)$ (equation (12)) were to be replaced by revenue net of labour costs, $\overline{p}x_A^s - \left(c^s + bx_A^s\right)x_A^s$ where $\overline{p}$ is an exogenously given price, and c^S and *b* are technology coefficients, giving the level and slope of average costs. If c^S is state dependent, then this model is evidently identical to the one above, with parameter α^S replaced by parameter $\overline{p} - c^s$. Uncertainty – in either costs or demand – means that profits are higher if input decisions can be postponed. The argument of this section shows that it also generates convexity of profits with respect to the location of component suppliers, implying that this uncertainty gives rise to clustering.

7. Summarizing the models

In section 3 we presented a standard economic geography model with orthodox transportation costs but no role for timeliness. In that model, there is no incentive for suppliers to agglomerate with one of the *ex ante* identical assemblers; in fact, the contrary holds, with efficiency tending to lead to a 50-50 split of suppliers in each assembly location.

The models of sections 4, 5, and 6 give the opposite conclusion. In each model, orthodox transport costs are absent, but a value for timeliness is introduced. In section 4, the value of timeliness comes from reducing the chance of costly production delays. In section 5, uncertainty about which products will be demanded creates a value for timeliness, though in this model we show how holding inventories *may* substitute for timely delivery. Finally, the section 6 model shows how uncertainty about the level of demand creates a demand for timely delivery.

In each of these three models, the demand for timely delivery creates a convexity in profits as a function of the location of suppliers, which can be thought of as an increasing marginal value of timeliness. Because of this convexity, there is in each model a force for agglomeration: all the suppliers will tend to locate in the same region as one of the *ex ante* identical assemblers, giving that assembler full benefit of flexibility while the other assembler makes due with non-timely delivery.

8. Policy implications

Governments are perennially interested in regional economic development, and subsidies have often been used (and even more often proposed) as a means of sustaining regional economies. In particular, subsidies to manufacturing assembly plants have been justified in the hope that their presence in a region will trigger agglomerations of related activities. The benchmark model of section 3 offers some theoretical support for such a subsidy: starting from a world with one assembly plant with all suppliers located nearby, establishment of a second assembly plant elsewhere creates an incentive for some suppliers to move near the new plant. This is because of the decreasing marginal value of proximity in such a model: the first supplier that moves to the location of the new assembler will generate greater value as a result.

In contrast, our models of timeliness deliver the opposite conclusion. Because of the increasing marginal value of timeliness (and hence proximity), there is no incentive for any supplier to move to the location of a new assembly plant. If these models apply, we would expect new assembly plants that locate far from existing plants (for whatever reason) to not be followed by their suppliers. As shown by Klier (1999), this is what has happened in the US auto industry: assembly plants established far from the "auto corridor" as a result of government subsidies (BMW in South Carolina, Mercedes Benz in Alabama) or private incentives (NUMMI in California) have not been followed by a substantial number of suppliers.[10]

9. Concluding comments.

Just-in-time production methods have been researched extensively in the management literature, but have received almost no attention in economics.[11] This paper has taken a step towards redressing this imbalance. In an uncertain environment the benefits of securing timely delivery of components alters the efficient spatial organization of production. In a situation in which conventionally modeled monetary trade costs would lead to dispersed location of component suppliers, delay or uncertainty in delivery times cause clustering. The efficient organization of production requires the concentration of all component plants next to just one of several assembly plants.

[10] The "auto corridor" is the region in the middle of the United States where most auto production is concentrated. It includes seven contiguous states: Michigan, Ohio, Indiana, Illinois, Wisconsin, Kentucky, and Tennessee.

This a new mechanism for agglomeration. In this paper we have developed the idea in a simple framework in which the final product is non-tradable, it is prohibitively costly to have multiple plants producing the same component, and factor prices are fixed. Each of these assumptions could be relaxed, embedding the mechanism in a wider economic environment.

[11] See Cremer (1995) for a rare exception.

Appendix

Parameter values used in figures

Figure 1: $r_A = r_B = 1$, $\sigma = 5$, $p = 1.5$, $\tau = 1.5$, $N = 10$.

Figure 2: $v_A^0 = 1$, $\delta p + \beta w_A = 1.5$, $q = 0.9$.

Figure 4: $r_A = r_B = 1$, $\mu = 0.25$, $\gamma = 0.03$, $\bar{p} = 1$, $\phi(k)$ CES with $\sigma = \delta = 0.33$.

Figure 5: $r_A = r_B = 1$, $\rho = 0.45$, $\alpha^H = 20$, $\alpha^L = 12$, $\beta = 1$

Section 5: CES preferences over component varieties: Let the price of the final product be equal to consumer utility, taking the form $\bar{p}\phi(k)$ where

$$\phi(k) = \left[\left(\frac{N-k}{N}\right) + \left(\frac{k}{N}\right)\delta^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$$

where the right-hand side is a CES function over the N components, proportion $(N - k)/N$ of which are perfect, and proportion k/N mismatched. $\bar{p}$ is the value of a perfectly matched component, and $\delta < 1$ is the utility loss for each mismatched component i.e. the loss if a component is a low preference variety rather than a high preference variety. Clearly, $\phi(0)=1$ and $\phi(N) = \delta$. $\phi(k)$ is decreasing in k (since $\delta < 1$) and is convex providing σ is finite. First and second derivatives are

$$\frac{d\phi(k)}{dk} = \frac{\sigma}{\sigma-1}\left(\frac{\delta^{\frac{\sigma-1}{\sigma}} - 1}{N}\right)\left[\left(\frac{N-k}{N}\right) + \left(\frac{k}{N}\right)\delta^{\frac{\sigma-1}{\sigma}}\right]^{\frac{1}{\sigma-1}} < 0$$

$$\frac{d^2\phi(k)}{dk^2} = \frac{\sigma}{(\sigma-1)^2}\left(\frac{\delta^{\frac{\sigma-1}{\sigma}} - 1}{N}\right)^2\left[\left(\frac{N-k}{N}\right) + \left(\frac{k}{N}\right)\delta^{\frac{\sigma-1}{\sigma}}\right]^{\frac{2-\sigma}{\sigma-1}} > 0$$

The condition $\bar{p}\mu\left[1 - \phi(N)\right] > \gamma N$ becomes simply $\bar{p}\mu\left[1 - \delta\right] > \gamma N$.

Section 6: Derivation of equation (19).

Total differentiation of (15) gives

$$\frac{dV_A}{dN_B} = \rho \frac{dv_A^H}{dN_B} + (1-\rho)\frac{dv_A^L}{dN_B} - N_B r \frac{dx_B}{dN_B} - rx_B.$$

By (12) and (14), this is

$$\frac{dV_A}{dN_B} = \rho\lambda^H \frac{dx_A^H}{dN_B} + (1-\rho)\lambda^L \frac{dx_A^L}{dN_B} + \rho r x_A^H + (1-\rho) r x_A^L - N_B r \frac{dx_B}{dN_B} - rx_B.$$

Using values of λ and of x_A^L and x^B from (17) gives equation (19) of the text.

References

Binmore, K.G. and P. Dasgupta, (1987) “Nash Bargaining III” in *The Economics of Bargaining* K.G. Binmore and P. Dasgupta (eds), Blackwell, Oxford.

Cremer, J. (1995) “Towards an economic theory of incentives in just-in-time manufacturing” *European Economic Review*, 39, 432-439.

Dicken, P. (1998) *Global shift; transforming the world economy*, Chapmans, London.

Evans, C and J. Harrigan (2003), “Distance, time, and specialization”, NBER Working Paper no. 9729 (May)..

Fujita, M. P. Krugman and A.J. Venables (1999), *The spatial economy: cities, region and international trade*, MIT Press: Cambridge MA

Fujita, M. and J-F Thisse (2001) *The economics of agglomeration; cities, industrial location and regional growth*, Cambridge University Press: Cambridge UK.

Hummels, D. (2001), ‘Time as a trade barrier’, mimeo Purdue University.

Jones, G.R., J.M. George and C.W.L. Hill (2000), *Contemporary management*, McGraw Hill, Boston.

Klier, Thomas, 1999, “Agglomeration in the U.S. auto supplier industry”, Federal Reserve Bank of Chicago *Economic Perspectives*, issue Q I, pages 18-34.

Kremer, M. (1993), “The O-ring theory of economic development”, *Quarterly Journal of Economics*, 108, 3, 551-575.

Rosenthal, Stuart S. and William C. Strange (2001). “The Determinants of Agglomeration”, *Journal of Urban Economics* **50**, 191-229.

Rosenthal, Stuart S. and William C. Strange (2003). “Evidence on the Nature and Sources of Agglomeration Economics” in *Handbook of Urban and Regional Economics*, eds J.V. Henderson and J-F Thisse, forthcoming.

Storper, M. and E. Leamer (2001) "The economic geography of the internet age", NBER Working Paper 8450.

Storper, M. and A.J. Venables (2003) "Buzz; face to face contact and the urban economy", Centre for Economic Performance discussion paper 598, and forthcoming *Journal of Economic Geography*.

Sutton, J. (1986), "Non-cooperative bargaining theory; an introduction", *Review of Economic Studies*, LIII, 709-724.

Venables A.J. (2001) "Geography and international inequalities: the impact of new technologies" in *Annual World Bank Conference on Development Economics* 2001/2, eds B. Pleskovic and N.H. Stern.

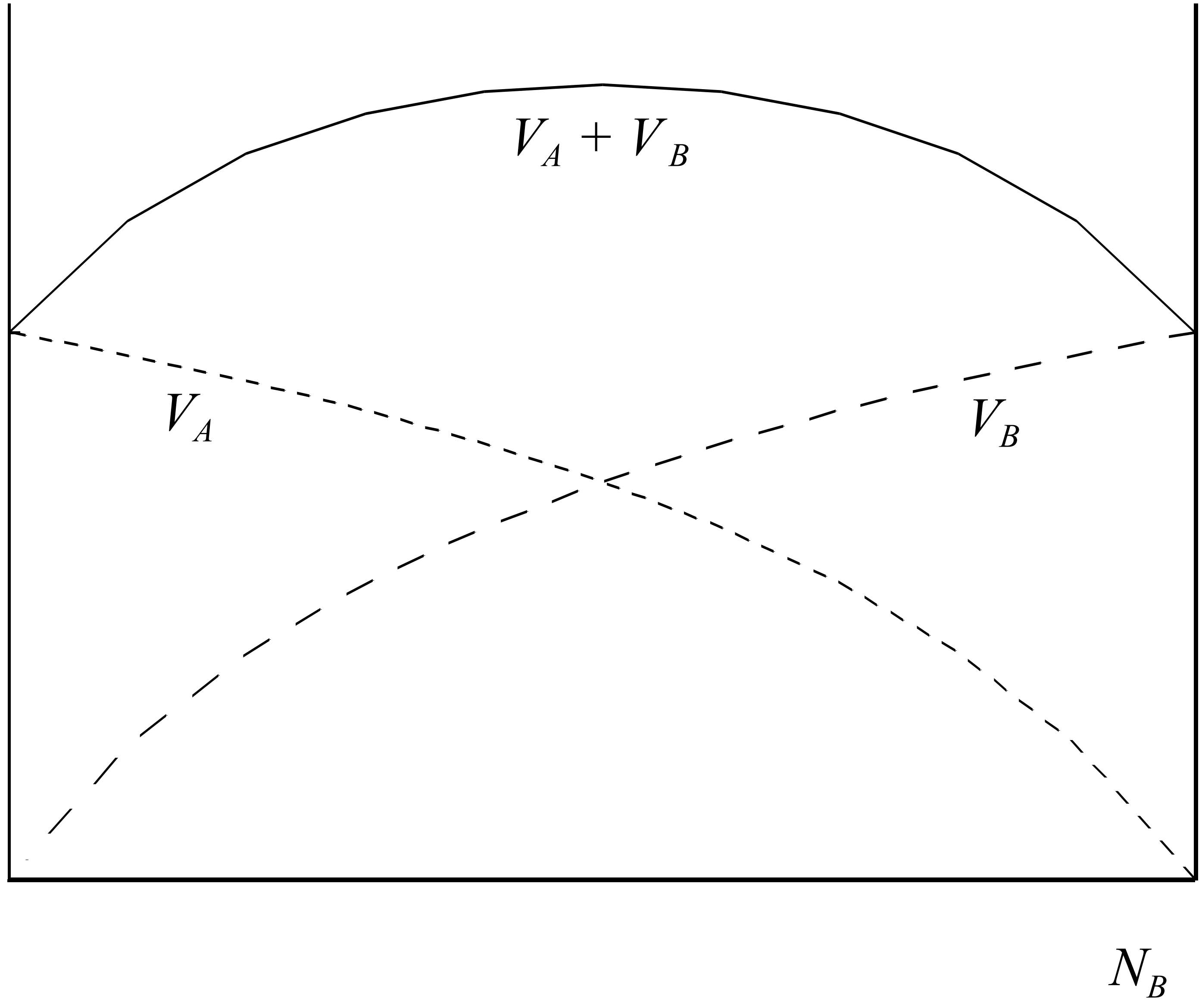


Figure 1: CES assembly (σ = 5)

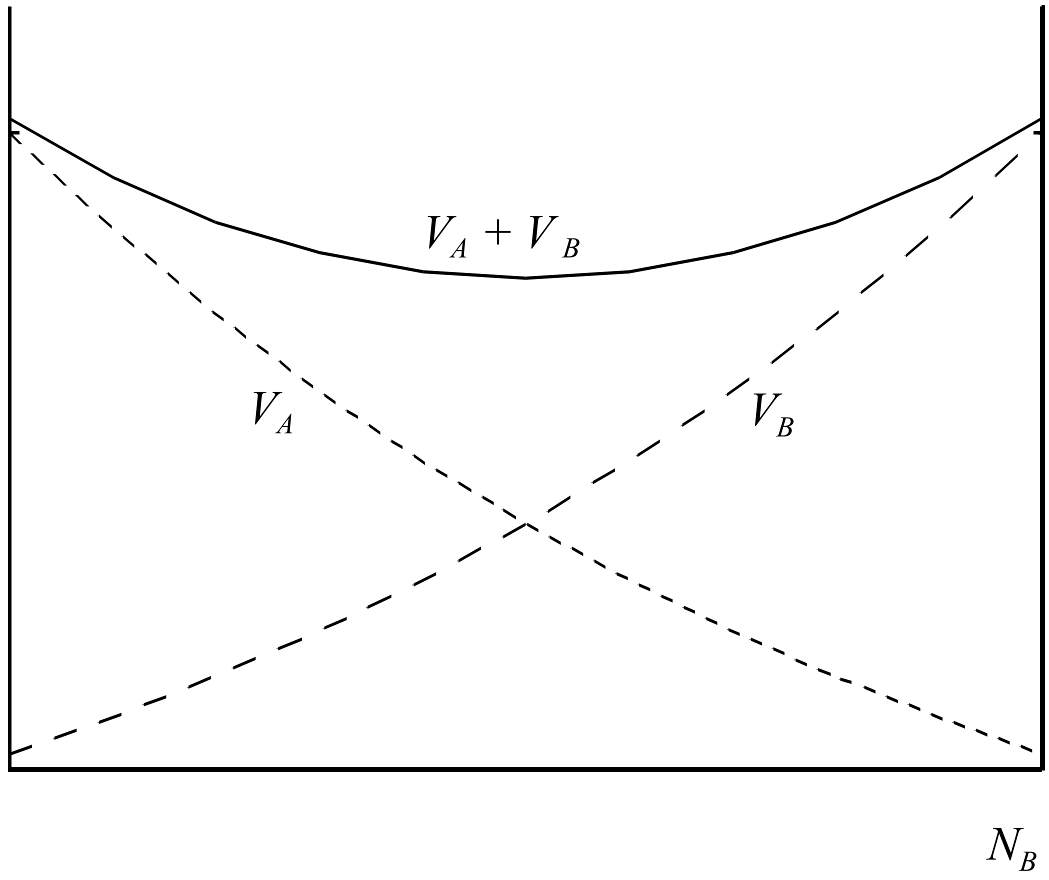


Figure 2: Arrival uncertainty

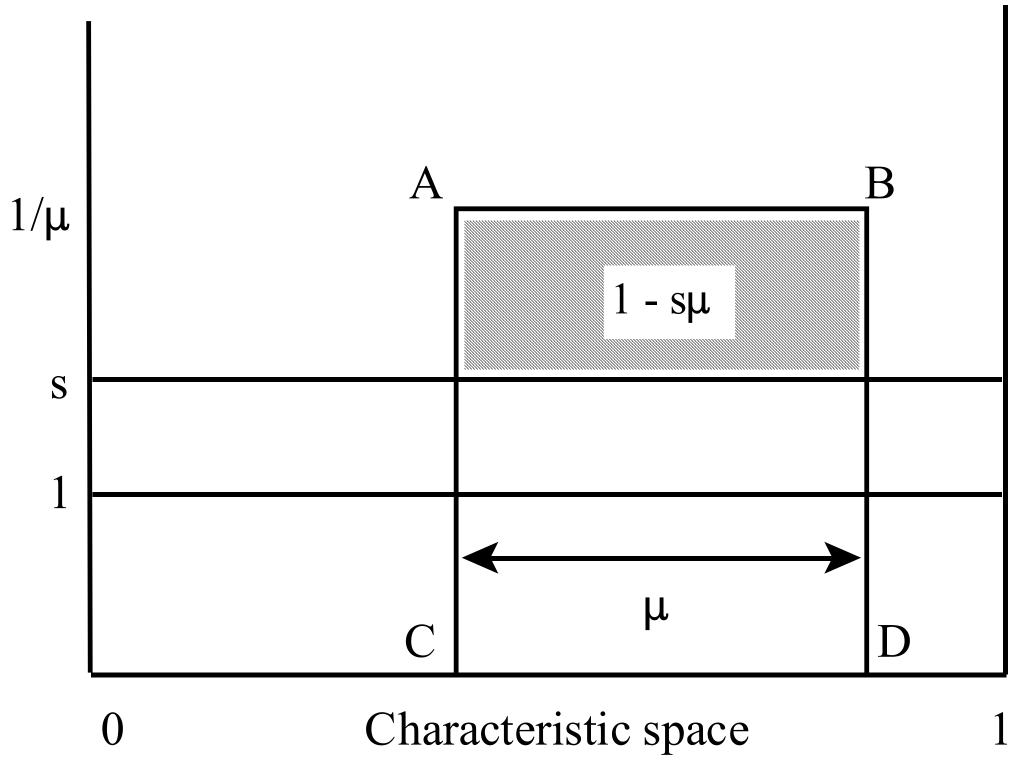


Figure 3: Demand for characteristics

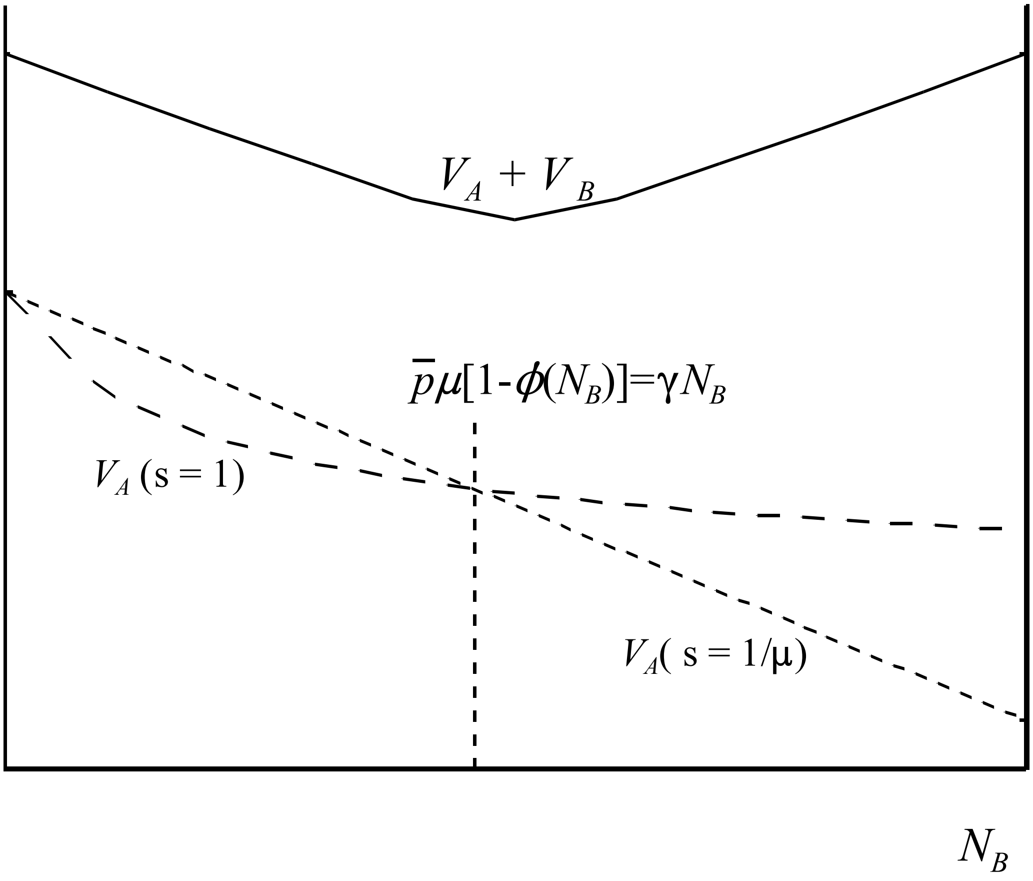


Figure 4: Inventory choice

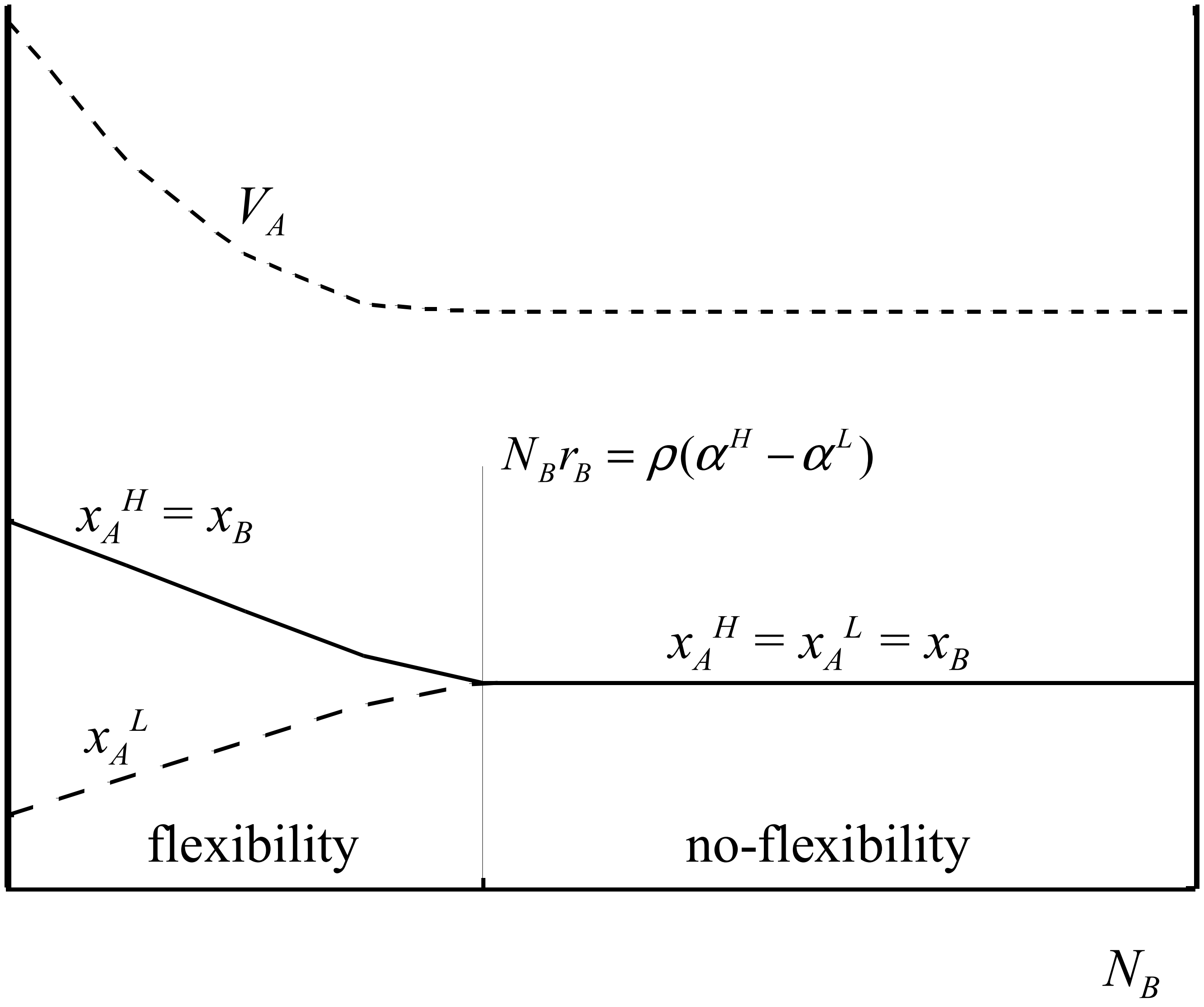


Figure 5: Demand level uncertainty