Board of Governors of the Federal Reserve System

International Finance Discussion Papers

Number 766

May 2003

Distance, Time, and Specialization

Carolyn Evans and James Harrigan

Distance, Time, and Specialization

Carolyn Evans and James Harrigan[*]

Abstract: Time is money, and distance matters. We model the interaction of these truisms, and show the implications for global specialization and trade: products where timely delivery is important will be produced near the source of final demand, where wages will be higher as a result. In the model, timely delivery is important because it allows retailers to respond to fluctuating final demand without holding costly inventories, and timely delivery is only possible from nearby locations. Using a unique dataset that allows us to measure the retail demand for timely delivery, we show that the sources of US apparel imports have shifted in the way predicted by the model, with products where timeliness matters increasingly imported from nearby countries.

Keywords: international trade, transportation costs, apparel.
JEL Classification: F1

[*] Evans: International Finance Division, Board of Governors of the Federal Reserve, Washington, D.C. 20551, <carolyn.l.evans@frb.gov>. Harrigan: International Research Department, Federal Reserve Bank of New York, 33 Liberty Street, New York, NY 10045, <james.harrigan@ny.frb.org>. We thank Frederick Abernathy, John Dunlop, Donald Foote, Janice Hammond, and David Weil for help with the data. Raymond Guiteras, Stefan Papaiannou, Rohit Vanjani and Irina Telyukova provided excellent research assistance. We also thank audiences at the NBER, CEPR, LSE, UCLA, UCSC, Syracuse, and Florida. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System, the Federal Reserve Bank of New York, or of any other person associated with the Federal Reserve System.

1 Introduction

If you want to sell something, it helps to be close to your customers. This truism leaves unanswered the question *why* proximity to one's market is good for business. The answer given by economic geography is very simple: transportation costs are increasing in distance, so it is more costly to deliver products to consumers far away than to those near the production location. With mobile factors of production, this "market access" motive leads to agglomeration near the source of final demand. When factors cannot move, remote factors will be paid less in equilibrium than those that are fortunate enough to be located near their customers. Redding and Venables (2001) provide strong evidence for this inequality effect of distance.

Another truism is "time is money". This explains why some goods are shipped by air, even though surface transport is invariably cheaper: customers are often willing to pay a substantial premium not to have to wait for their ship to come in. As documented by Hummels (2001), the premium that must be paid for air shipment far exceeds the interest cost savings on inventory in transit. This strong urge to save time, even at great expense, implies a powerful force for agglomeration and/or spatial inequality that is distinct from the transport-cost-economizing motive emphasized in the economic geography literature.

This paper studies the interaction of these two truisms. We present a simple model of the demand for timeliness and its implications for international specialization and trade. The model is motivated by the experience of the global apparel and textile industry, which saw two simultaneous trends in the 1990s. The first was the rise of "lean retailing", a set of business practices made possible by advances in information technology that allow retailers to hold small inventories and respond rapidly to fluctuations in consumer demand. The second trend was a shift in the location of production away from lower wage locations in Asia toward higher-wage locations in Mexico and the Caribbean.[1] We argue that these trends are related: lean retailing creates a demand for timeliness, which can only be met by producers located near the US market. Our model shows that the result of an increased demand for timeliness is that wages will be higher in locations near the source of final demand, with lower wage distant locations specializing in products where timeliness is less important. This economic geography result

[1] Monthly manufacturing wages in Mexico were three times as high as in China in 1998 (ILO website).

comes from a model with traditional transport costs of zero and constant returns to scale in production.

An implication of our model is that the shift in the sourcing of US apparel toward Mexico and the Caribbean is disproportionately concentrated in goods where timeliness is important. We test this implication on a unique dataset that combines product-level information from a major department store chain with detailed information on trade flows and trade barriers. We find strong evidence that nearby producers are increasingly specialized in goods where timeliness is important to retailers, as predicted by the theory.

Our paper is one of a very few that study the importance of timeliness in determining trade patterns, and the first to build careful microfoundations for the demand for timeliness and provide empirical evidence on its importance. Deardorff (2002) considers some of the same theoretical issues that we do in an insightful but informal way. His conclusion that time-sensitive goods will be produced by capital-intensive countries is a consequence of his assumption that speed is capital-intensive. Deardorff also conjectures that remote countries are less likely to specialize in time-intensive products, a result that we establish theoretically and empirically. Venables (2001) discusses the tradeoff between proximity and production costs, and argues that technological change that makes timely production easier will lead to production shifting closer to the center, a result for which we provide microfoundations below.

2 Flexible production and lean retailing

Selling clothing is a nerve-wracking business. Consumer tastes are volatile, and retailers are haunted by the prospect of having to liquidate vast inventories of unpopular clothing at the end of a selling season and, equally painfully, of running short of suddenly popular styles. "Lean retailing," the combination of low inventories and frequent restocking, offers a partial solution to these problems.[2] With low inventories, stores will not be stuck with large amounts of unsold goods even if demand collapses. With frequent restocking, stores will not run short of popular items. Lean retailing requires

1. *Bar codes,* that allow retailers to keep daily track of sales of each of the tens of thousands of products that they stock.

2. *Electronic data interchange*, which is a system of linked computer networks that make it possible for retailers to communicate quickly and cheaply with suppliers.

[2] The discussion here is drawn from Abernathy et al (1999).

3. *Modern distribution centers* that rapidly channel goods from suppliers to sales locations.

The essence of lean retailing is to respond rapidly to demand fluctuations instead of holding large inventories. The final link is that production cannot be too far from the sales location, because goods need to be moved quickly. This demand for timeliness leads to a demand for proximity, since shipping time is increasing in distance.

Demand variability is not the same in all categories of apparel, of course. Some items have very predictable demand, so that lean retailing offers little benefit. Other goods have demand that is so unpredictable, and which have such short selling seasons, that the classic inventory problem cannot be mitigated even with deft application of lean retailing strategies. In the middle are goods that have variable demand, and selling seasons that are long enough that it is feasible to replenish supplies if demand conditions warrant. In the jargon of retail management, goods that are ordered more than once per selling season are called “replenishment” goods, while goods that are ordered only once per season are "non-replenishment" items.

From the standpoint of producers, lean retailing demands great flexibility. If they want to sell to lean retailers, producers must be able to adjust output rapidly and ship products quickly. The benefit for flexible producers is that they can charge a premium over their non-flexible competitors, who can only compete on selling cost and not on timeliness.

We now turn to the implications of lean retailing for the equilibrium location of production and international wage differentials.

3 Timeliness in general equilibrium: a model

The purpose of our model is to derive the equilibrium pattern of specialization and wages in a world where flexible production is only possible if production is located near the source of final demand. This assumption comes from the more primitive assumption that distant production locations are sufficiently far away that shipping times are too long to meet the deadlines required by lean retailers. It is helpful to keep a stylized geography in mind, with the United States being the source of final demand, the Caribbean/Mexico being adjacent to the US and hence close enough to engage in flexible production, and Asia located so far away that flexible production is impossible.

We build the model in four stages. First, we derive optimal production for flexible and non-flexible firms separately, where each risk-neutral firm sells a unique product.[3] If all firms have identical costs, then flexible firms will make greater profits, by the convexity of the profit function. This means that locations where flexible production is impossible (Asia) can only compete if they offer wages lower than those in locations where flexibility is possible (the Caribbean). Second, we consider the tradeoff between flexibility and costs, and derive an expression that shows that the number of firms locating in Asia is an increasing function of the Caribbean's wage premium. Third, we derive the labor market equilibrium conditions for the two regions for a given international division of labor. Finally, we use the locational and labor market equilibrium conditions to solve for the equilibrium wages and pattern of specialization.

3.1 The firm's production decision

The structure of the model is driven by demand. In each year, demand is realized twice, and all firms have to make production decisions before the first period. Firms with production locations nearby have the option of producing again after the first period, while firms with faraway plants do not have this flexibility. For now, we take the location of firms as given. Inventory can be carried over at no cost between periods within a year, but any inventory unsold at the end of the year has a zero price in subsequent years.

An individual firm faces a linear inverse demand for its product in each period, given by

$$p = a - b \cdot s \tag{1}$$

where p is price, s is sales, and a and b are parameters. The source of uncertainty in demand comes from fluctuations in the intercept a:

$$a \in \{a_L, a_H\} \tag{2}$$

where $a_H > a_L$, and the average value of a is $\overline{a}$. The production function is as simple as possible - output equals labor input, so that marginal cost is just equal to the wage w. Setting marginal revenue = marginal cost gives the firm's per period optimal sales as

$$s = \frac{a - w}{2b} \tag{3}$$

[3] For our purposes, there is no benefit to modeling a separate retail sector, so we assume that producers sell directly to consumers.

The firm would obviously prefer to wait until demand is realized before it decides what to produce. By assumption, all firms have to produce before first period demand is realized, while only flexible firms can produce between periods.

In the appendix, we work out the details of optimal production plans for risk-neutral firms, but the solution is intuitive. Non-flexible firms simply produce twice expected optimal sales:

$$q_1^N = 2\frac{\overline{a}-w}{2b} = 2q^* \tag{4}$$

where q is output, subscripts denote period production and the superscript N identifies non-flexible firms, and q^* is just ex-ante optimal sales in each period. Flexible firms will produce enough in the first period to sell the optimal amount if demand is high:

$$q_1^F = \frac{a_H - w}{2b} \tag{5}$$

If demand turns out to be low, they will sell the optimal amount given low demand, and hold inventory into the next period. We can summarize the flexible firm's actions, in the order in which decisions are made, as

$$q_1^F = \frac{a_H - w}{2b}$$

$$s_1^F = q^* + \frac{a_1 - \overline{a}}{2b}$$

$$q_2^F = q^* - \frac{a_H - a_1}{2b}$$

$$s_2^F = q^*$$

The ex-post optimal amount to sell in the first period, s_1^F, depends on the realization of a_1 in the first period, and the firm's period 2 production just offsets the demand surprise in period 1.

Actual sales by non-flexible firms are

$$s_1^N = q^* + \frac{a_1 - \overline{a}}{4b}$$

$$s_2^N = q^* - \frac{a_1 - \overline{a}}{4b}$$

Note that first period sales by non-flexible firms respond less to demand shocks than do the sales of flexible firms. Essentially, non-flexible firms hedge: if demand is high in period 1, they sell less than if they were flexible, because they want to make sure they have enough output to sell if demand is high in the second period. Similarly, if demand is low in period 1, the firm sells more than it would if it were flexible, because it doesn't want to be stuck with huge inventories if demand is low again in the second period.

This analysis illustrates that even if demand is uncorrelated across periods (as we assume for simplicity), flexible firms will always choose to produce twice, so that they can take advantage of what happens in period 1. Average output and sales for the two types of firms are the same, but output is more variable for the flexible firm, and it is this variation that leads to higher average profits for the more flexible firm.[4]

3.2 The firm's location decision

We now address the question of where firms locate. Clearly, if costs were the same all firms would like to be flexible, and the demand for labor in the non-flexible location, Asia (*A*), would be zero. If wages in the flexible location, the Caribbean (*C*), are higher than in *A*, so that $\hat{w} = w_C - w_A > 0$, then firms face a tradeoff between the benefits of flexibility and the costs of paying higher wages. If demand is very variable it may be worth paying the higher wages to get the benefits of flexibility; but if demand is not very variable or if the wage differential is large, firms will choose the non-flexible location.

For firms wishing to benefit from flexibility, an alternative to locating in *C* is to locate in *A* and to ship goods by air instead of by ship. But airfreight is expensive: for example, US importers paid a premium for air over surface shipment that averaged 25% of the transported goods value in 1998 (Hummels, 2001). To keep the focus on the trade-off between wage costs and flexibility, we assume that the cost of airfreight exceeds the equilibrium wage savings from producing in *A*.

While the degree of demand variability affects a firm's desire for flexibility, the length of a product's selling season affects whether flexible production is technically feasible. Some products (such as New Year's Eve gowns) have a very short selling season, which makes reordering once initial demand is realized impractical. Other products (such as men's white cotton underwear) are sold year round, so that there is plenty of time to reorder once initial

[4] Note that a_2 has no effect on second period sales for either type of firm - s_2 is predetermined once a_1 is realized. Nonetheless uncertainty in period 2 does have an effect on the solution of the non-flexible firm's problem: first period sales would respond more to a_1 if there were no uncertainty about a_2.

demand is realized. Most products having selling seasons in between these extremes. In terms of our model, the calendar time between periods might be two weeks for New Year's Eve gowns and four months for men's white cotton underwear.

We now suppose that there is a continuum of monopolists indexed by *i* on [0,1], all with identical cost functions and facing similar demand curves for the products that they produce:

$$p(i) = a(i) - bs(i) \tag{6}$$

What distinguishes products from each other is length of the selling season and variability of demand. Only a subset of goods have selling seasons long enough that flexible production is technically feasible; of these, only some have demand that is variable enough to make flexible production the profit maximizing strategy. We order goods so that products in [0, i_u) have long enough selling seasons for flexibility to be feasible, where $i_u \leq 1$ is a parameter. Firms located in $i \in [i_U, 1]$ are technologically incapable of engaging in flexible production.

For the products in [0, i_u), we order them so that variance in demand is increasing in *i*. In particular, we suppose that the variance of *a* is proportional to *i*,

$$V\left[a(i)\right] = i \cdot \sigma^2 , \qquad i \in [0, i_U) \tag{7}$$

where σ^2 is a parameter.[5]

Firms that are technologically capable of flexible production, $i \in [0, i_U)$, will choose the location that maximizes expected profits, trading off the benefits of flexibility with the higher wages that must be paid to produce in *C*. For a given $\hat{w}$, firms with the least variable demand will choose to locate in *A*, while firms with more variable demand will choose to produce in *C*. The marginal firm, located at i_L, is just indifferent between producing in *A* or *C*, and this indifference defines a relationship between i_L and $\hat{w}$: the higher is $\hat{w}$ the higher will be i_L, since fewer firms will find it worthwhile to pay the higher wages necessary to produce in *C*. We show in the appendix that this relationship is given by

$$i_L = \frac{8\overline{a}\left(w_C - w_A\right) - 4\left(w_C - w_A\right)\left(w_C + w_A\right)}{\sigma^2} \tag{8}$$

We call this relationship the *QQ* curve, and it is graphed in Figure 1.[6] Note that so far we have not ruled out the possibility that $\hat{w}$ could be so high that $i_L > i_U$, which would imply that no

[5] This proportionality assumption simply makes the algebra easier - all that is required for the model is that there is a monotonic relationship between *i* and *V*[*a(i)*]..

firms want to locate in the high-cost Caribbean. We also don't *a priori* rule out $\hat{w} < 0$, but if this were the case no firms would want to locate in *A*.

3.3 Labor market equilibrium

We now turn to the labor market in each potential production location. In the background is a Ricardian international trade model, where the United States has a comparative disadvantage in apparel relative to *C* and *A*, who have identical technology. We can pick parameter values and country sizes to guarantee that we are in a complete specialization equilibrium, where the US produces only its' comparative advantage good (call it machinery) and *A* and *C* produce only apparel. Machinery will be our numeraire.

Aggregate labor supplies *L* in *A* and *C* are fixed. The average flexible producer has per-period labor demand equal to average output q^*, so that total annual labor demand per average flexible firm is $2q^*$. Each non-flexible firm has the same labor demand, so total demand from each non-flexible producer is also $2q^*$.

The number of flexible firms is given by the distance between the lower and upper bounds i_L and i_U. As long as $i_U > i_L$ this distance is just $i_U - i_L$. Substituting for q^* and setting the demand for Caribbean labor equal to the fixed supply gives the labor market clearing condition for *C* as

$$w_C = \bar{a} - b\frac{L_C}{i_U - i_L} \tag{9}$$

If C were large enough to satisfy all the demand for labor by potentially flexible firms at a zero wage then the model breaks down, so we assume that C is small enough that this doesn't happen. By setting $i_L = 0$ we determine that this parameter restriction is

$$L_C < i_U \frac{\bar{a}}{b} \tag{10}$$

The remainder of the firms are inflexible and produce in A where wages are cheaper.[7] The corresponding labor market equilibrium condition is

[6] Equation (8) is a quadratic in w_A and w_C separately, which defines a three dimensional surface in i_L-w_A-w_C space. In the appendix, we show that the *QQ* curve is the locus of equilibrium wage differentials as a function of i_L.

[7] Firms $i \in [0, i_L)$ choose to produce in A because it is more profitable than producing in C, while firms $i \in (i_U, 1]$ produce in A because their selling seasons are too short for flexible production to be feasible.

$$w_A = \overline{a} - b\frac{L_A}{1+i_L - i_U} \tag{11}$$

Subtracting w_A from w_C gives an expression for the equilibrium wage differential as a function of country size and the international pattern of specialization:

$$\hat{w} = b\left(\frac{L_A}{1+i_L - i_U} - \frac{L_C}{i_U - i_L}\right) \tag{12}$$

For a technologically-fixed upper bound i_U, this relationship is convex and decreasing in i_L: the larger the share of potentially flexible production that goes to *A*, the lower is the wage differential between *A* and *C*. We call this the *LL* curve, and it is illustrated in Figure 1.[8]

3.4 General equilibrium

Putting the *QQ* and *LL* curves together gives our equilibrium, which is illustrated in Figure 1. Wages are higher in *C* than in *A*, and as a consequence some firms that are technologically capable of flexible production forgo that possibility in favor of the cheap wages available in *A*. Other firms, who face greater demand variability, find it worthwhile to pay the higher wages needed to produce in *C*.

As drawn the *LL* curve crosses the horizontal axis in the range (0, i_U), which guarantees that the equilibrium $i_L \in (0, i_U)$ and therefore $\hat{w} > 0$. This is only guaranteed if *C* is sufficiently small relative to *A*[9]:

$$\frac{L_C}{L_A} < \frac{i_U}{1 - i_U} \tag{13}$$

If this restriction is not satisfied, then $\hat{w} = 0$ and $i_L = 0$. In this case, all of the potentially flexible producers in the range $[0, i_U)$ and at least some of the firms in the range $[i_U, 1]$ will produce in *C*. Note that this restriction is automatically satisfied as i_U approaches 1: since all firms value flexibility in the limit, no firm will be willing to produce in *A* unless wages are lower there.

So far we have concentrated on the novel parts of our model, the production decisions and the determination of $\hat{w}$. It is straightforward but uninteresting to close the model, so we simply sketch the solution here. In order to generate a perfect specialization equilibrium, we

[8] We verify in the appendix that the *LL* curve is convex and asymptotically approaches i_U.

[9] to derive this inequality, set $\hat{w} = 0$ and solve (12) for i_L; imposing $i_L > 0$ then gives the condition. Note that we now have two restrictions on the size of C: it can't be too big either absolutely or relative to A.

assume that unit labor requirements for apparel in all countries are equal to unity. In *A* and *C*, the unit labor requirement for the numeraire is also unity, while it is less than one in the US. Residents of *A* and *C* have no taste for apparel, consume only the numeraire good*,* and have income only from labor. In a perfect specialization equilibrium, then, citizens in *A* and *C* simply trade their labor income for imports of the numeraire good. Gains from trade follow immediately.

Note, however, that the law of comparative advantage does not fully predict trade patterns in this model. Since they have identical preferences and technology, countries *A* and *C* have identical autarky prices. Not surprisingly, they do not trade with each other in equilibrium. What is surprising is that they export disjoint sets of products, and *A* gains less from trade than does *C* (since $\hat{w}$ = 0 in autarky and $\hat{w} > 0$ with trade). This is because geography is irrelevant in autarky but not when trade is possible (see Deardorff (2001) for another example of this theoretical phenomenon). The breakdown of comparative advantage has nothing to do with increasing returns, which are absent in the model. Nor (unlike Deardorff's model) is it due to transportation costs, which are zero here. The reason geography matters in our model is that shipping takes time, which makes proximity valuable even though the cost of shipping (in the usual sense of a charge for moving goods) is zero. Introducing shipping costs that increase with distance into our model would accentuate the equilibrium wage differential $\hat{w}$, but would not alter the conclusion that nearby countries specialize in goods where timely delivery is relatively valuable.

3.5 The spread of flexible production in general equilibrium

Abernathy et al (1999) make it clear that lean retailing spread slowly through the apparel sector during the 1990s. As technology improved and as management techniques diffused, more apparel firms became capable of producing flexibly. We model this as an exogenous increase in i_U, with the newly capable firms located in the interval Δi_U. With an increase in i_U, there are two possibilities:

1. all the products in Δi_U have variance less than $V\left(a(i_L^0)\right)$,
2. at least some of the new products in Δi_U goods have variance greater than $V\left(a(i_L^0)\right)$.

In the first case, there is no change in the equilibrium: the products in Δi_U were produced in *A* before and they still are. Even though it is now technically feasible for these products to be produced flexibly, it is not profitable to do so, so they stay in *A* where wages are low.

In the second case, products $j \in \Delta i_U$ such that $V[a(j)] > V\left(a(i_L^0)\right)$ can be produced more profitably in *C* than in *A* at the initial relative wage. This leads to a shift in labor demand away from *A* toward *C*, and the consequences are illustrated in Figure 2.[10] The *LL* curve shifts to the left, and i_L also shifts left (to i_L^1) but by less than Δi_U. As a result, wages rise in *C* relative to *A*, and the total number of firms producing in *C* increases. This story matches the account given in Abernathy et al (1999, Chapter 13): as more retailers adopted "lean retailing" strategies during the 1990s in response to diffusion of technology and management practices throughout the industry, this was matched by a shift of apparel sourcing from the Far East to the Caribbean Basin. Interestingly, in our model some producers shift from *C* to *A* when i_U increases: these are firms who just found it worth paying high *C* wages before, but who (given the small value they attach to flexibility) are now priced out of *C*'s labor market.

The model of this section gives two key empirical predictions:

1. Products produced in high-wage locations near the source of final demand are those that are ordered by final sellers more than once per selling season. Apparel retailers call these "replenishment" goods. Goods produced in distant low-wage locations are non-replenishment items.
2. As information technology improves and spreads, making flexible production feasible for a wider range of goods, it will cause shifts in the global pattern of trade and income. Countries closer to large sources of final demand will benefit at the expense of more remote locations.

This second prediction is particularly interesting for two reasons. First, it is a result about economic geography that comes from a model with no transport costs, no increasing returns, and no Dixit-Stiglitz preferences. In this regard, our model is similar to a von Thünen central place model, with the relatively transport-intensive goods locating near the exogenously given center, and wages declining with distance from the center, but our mechanism is wholly different. Second, it turns predictions about the "death of distance" on their head: in our model,

[10] We draw Figure 2 using the simplifying assumption that $V[a(j)] > V\left(a(i_L^0)\right)\ \forall\, j \in \Delta i_u$. This makes drawing the figure neater but has no analytical consequences. See the appendix for the details.

improvements in communications technology make distance matter *more* for incomes and trade in equilibrium, not less.

4 Empirical Evidence

In evaluating our model, we focus on US imports of apparel. As Figure 3 shows, there has been a dramatic shift in the sourcing of US apparel imports, with Mexico and Caribbean countries gaining at the expense of countries in Asia, particularly China/Hong Kong. Our model gives one explanation for this shift, but there are at least two others that are potentially important: changes in comparative advantage and changes in trade policy.

4.1 Labor costs

Apparel is an unskilled-labor intensive traded good, and is often considered the archetypal footloose manufactured product, with capitalists scouring the globe for the lowest wages. On this view, what matters for competitive advantage in apparel is low wages. It is difficult to get comparable data on wages in apparel production around the world, and impossible to get productivity-adjusted wages. As a first step, Table 1 shows the relative wage in overall manufacturing for China and Mexico from 1991 to 1998. The table illustrates that China has much lower wages than Mexico, but the ratio shrank from nearly 9 at the beginning of the decade to just over 3 in 1998. However, most of the drop in Mexican relative wages occurred between 1991 and 1995, and has stayed fairly flat since then. This is inconsistent with the behavior of market shares seen in Figure 3, in which Mexico's share accelerated in mid-decade. Furthermore, wages remain much higher in levels in Mexico than in China. Our tentative conclusion is that falling Mexican relative wages may have contributed to Mexico's growing success in exporting apparel, but do not completely explain it.

4.2 Trade Policy

A second explanation for changing trade patterns is changes in trade policy. The dominant instrument of trade policy for textiles and apparel is the Multi Fiber Arrangement or MFA, a Byzantine system of bilateral product specific quotas that dates back to the 1950s and which is very slowly being phased out.[11] No analysis of apparel trade can be credible without accounting for the MFA, so we do just that.

The MFA is extremely opaque, and to our knowledge we are the first researchers to assemble a comprehensive product-level time series on the US MFA program, which is

[11] It is due to disappear completely in 2005.

administered by a division of the Commerce Department called the Office of Textiles and Apparel or OTEXA.[12] Quota levels vary by product, year, and trading partner. We obtained records on the levels of all apparel quotas from 1990 to 1998, along with the "fill rate", which is the percentage of the quota used. OTEXA uses their own import classification system to administer the MFA, which has no simple relationship to any other US or international system of reporting trade data.[13] The product categories are broken down by type of fiber (cotton, wool, silk, man-made, and other), and are fairly broad: categories include "dresses," "sweaters," "underwear," and the like.

Our trade data on apparel imports, tariffs, and transport costs come from CD-ROMS purchased from the US Commerce Department. This data is reported at the 10-digit HS level, which is the finest level of disaggregation available. Among other things, the data includes information on import values, import quantities, tariffs, transport costs, and source country. In analyzing the data, we aggregate up to the OTEXA import classification system.

Figure 4 summarizes the quota data. It shows a histogram of quota fill rates across all sources of apparel imports, weighted by import values.[14] If we define a binding quota as one with a fill rate of 90% or above, Figure 4 shows that about 40 percent of US apparel imports came in under binding quotas throughout the 1990s, and that there has been very little change in this proportion despite the liberalization promised under the Uruguay Round.

Tariffs also remain an important trade restriction for US apparel imports. Figure 5 shows the incidence of tariffs, and contrary to the quotas seen in Figure 4, there is clear evidence of liberalization: in 1990 and 1991, about half of US imports paid tariffs of over 16%, and virtually none came in duty-free. By 1998, high tariffs were much less prevalent, and about 20% entered nearly duty free (with tariffs of less than 2%).

The overall trends visible in Figures 4 and 5 obscure important variation across trading partners. Figures 6 and 7 illustrate the importance of two important trade policy initiatives, NAFTA and the Caribbean Basin Initiative (CBI). Figure 6 shows that in 1990, Mexico and the Caribbean faced tariffs similar to those faced by other US import sources. By 1998, Mexico had very privileged access, with virtually all apparel imports entering with at most nominal tariffs.

[12] OTEXA's website is fairly informative, and can be found at http://otexa.ita.doc.gov/default.htm

[13] The system is documented at http://otexa.ita.doc.gov/corr.stm

[14] Imports that are not subject to any quota at all can be thought of as facing an infinite quota and hence have a zero fill rate.

The Caribbean saw less dramatic changes over the decade, but clearly these countries' market access relative to all countries other than Mexico improved significantly.

Figures 8 and 9 show the regional evolution of MFA incidence. Interestingly, both Mexico and the Caribbean faced *more* binding quotas at the end of the decade than they did in 1990, perhaps reflecting a political economy response to rapid import growth from these regions (alternatively, unchanged quotas may have become binding as import demand grew). East Asia, China, and Hong Kong did not see major changes in their incidence of binding quotas, while South Asia saw a big increase: in 1990 less than 20 percent of South Asian imports entered under a binding quota, and this proportion almost quadrupled by 1998.

This eyeball analysis of trade policy strongly suggests that NAFTA and the CBI are at least partly responsible for the shifts in apparel import sourcing seen in Figure 3. The analysis also suggests that controlling for the effects of the MFA is crucial, since such a large share of apparel imports come in under binding quotas.

4.3 Product characteristics and trade: testing the demand for timeliness model

In this section we develop an empirical model that allows us to test a central implication of our model while taking account of other important determinants of apparel imports.

The model predicts that apparel products that are subject to rapid retail replenishment will be sourced from countries close to the US, where they can be imported quickly in response to changing demand conditions. We use a unique, proprietary data source from a major department store chain to identify such products. The chain has stores all across the country, and we have information on clothing sales at all of their stores in 2001, including which items are replenished and in what proportions. These replenishment proportions are aggregated across stores and product lines to give aggregate replenishment proportions by broad product category. Confidentiality precludes us from illustrating this data, but the range of replenishment across products is from 0 to 67%.

Our approach to testing the model is simple: we specify a reduced form equation for desired imports, and assume that actual imports are given by the minimum of desired imports and the exogenous import quota. With the notation

m_{ict} = log level of real (physical quantity) imports of product *i* from country *c* in year *t*,

m^*_{ict} = log unconstrained imports, and

q_{ict} = log quota level on product *i* in country *c*, we then have

$$m_{ict} = \text{Min}[\, m^{*}_{ict}\, ,\, q_{ict}\,] \qquad (14)$$

Unconstrained imports depend on country-time and product-time dummies, as well as ad-valorem trade resistance including tariffs and transport costs, given by τ_{ict}:

$$m^{*}_{ict} = \mu_{it} + \mu_{ct} + \alpha\tau_{ict} \qquad (15)$$

We assume that timeliness was irrelevant in period 1, because the development of lean retailing was in its infancy at the beginning of the 1990s. By the end of our sample in 1998, our model predicts that replenishment product categories will be sourced from countries near the US. We capture this with an interaction effect between replenishment proportion r_i and a dummy d_c for proximity to the US (equal to one for Mexico and the Caribbean countries). The level equation for unconstrained imports in each period becomes

$$\begin{aligned} m^{*}_{ic1} &= \mu_{i1} + \mu_{c1} + \alpha\tau_{ic1} \\ m^{*}_{ic2} &= \mu_{i2} + \mu_{c2} + \alpha\tau_{ic2} + \beta r_i d_c \end{aligned} \qquad (16)$$

Looking at import growth from period 1 to 2, there are four possible situations:

a. Quota binds in both periods $\Delta m_{ic} = \Delta q_{ic}$ (17a)

b. Quota slack in both periods $\Delta m_{ic} = \mu_i + \mu_c + \alpha\Delta\tau_{ic} + \beta r_i d_c$ (17b)

c. Quota binds in 1st period only $\Delta m_{ic} = \mu_{i2} + \mu_{c2} + \alpha\tau_{ic2} + \beta r_i d_c - q_{ic1}$ (17c)

d. Quota binds in 2nd period only $\Delta m_{ic} = \mu_{i1} + \mu_{c1} - \alpha\tau_{ic1} + q_{ic2}$ (17d)

Since our primary interest is in estimating the importance of timeliness on import growth, which is measured by β, observations in cases a and d are irrelevant. As it happens, there are only a very small number of observations in case *c*, so we focus on estimating the model using solely observations where imports were unconstrained in both periods, equation (17b) (where we define $\mu_i = \mu_{i2} - \mu_{i1}$, $\mu_c = \mu_{c2} - \mu_{c1}$).

While simple, the specification in (17b) controls for most of the factors that could affect import growth. The country dummies μ_c control for influences such as factor prices, the country's average level of tariffs and quota restrictiveness, and other country-specific effects. The product dummies μ_i account for the average rate of growth of imports in the category, as well as the average world level of tariffs and quotas on that product. We assume that, aside from trade costs and the timeliness effect, all other idiosyncratic influences on imports are orthogonal, and we summarize their effect in a residual error term. The timeliness effect β answers the question: do imports of high-replenishment goods grow more rapidly from Mexico and the Caribbean than they do from the rest of the world? Our model says the answer is yes, and predicts $\beta > 0$.

We estimate (17b) on a panel of apparel import growth across products and countries, over the period 1991 to 1998. Rather than look at year-to-year variation, we focus on total growth over the seven-year period. There are 3,177 observations in our full sample, of which 2,753 are not quota constrained and are therefore appropriate for estimating the regression model. Table 2 gives summary statistics for the full and unconstrained sample; our comments here will refer to the latter. A remarkable feature of the data is how skewed the distribution of import growth is: the median is a fast but reasonable 50%, while the mean is an outlandish 6,763%. This arises because for a substantial share of the observations, imports were extremely low in 1991 and large in 1998, so that many growth rates are very high (in fact, the 75th percentile of import growth is over 400%). Of course many other growth rates are negative, with the 25th percentile equal to -60%.

The change in trade frictions is less skewed, with a mean of -4.52 percentage points and a median of -2.87 percentage points, but there is a lot of variation (the standard deviation of the change in trade frictions is 14 percentage points). Most of the fall in trade frictions is due to declining tariffs rather than falling transport costs (-2.9 and -1.6 percentage points on average respectively). Three-quarters of all the changes in trade barriers were negative, reflecting the broad reductions in tariff barriers seen in Figure 5.

The extreme skewness of import growth suggests that an estimator that assumes a symmetrical distribution will be inefficient and probably misleading. We address this issue by defining "bounded" import growth as follows:

$$G_{ic} = 100 \cdot \frac{m_{ict} - m_{ic,t-1}}{0.5\left(m_{ict} + m_{ic,t-1}\right)} = 200 \cdot \frac{m_{ict} - m_{ic,t-1}}{\left(m_{ict} + m_{ic,t-1}\right)}$$

The ordinary measure of percentage growth, $g_{ic} = 100 \cdot \frac{(m_{ict} - m_{ict-1})}{m_{ict-1}}$, is the change divided by the initial level, while bounded growth is the change divided by an average of the beginning and ending period levels. As a result, this measure, used by Davis, Haltiwanger and Schuh (1996) in their studies of manufacturing plant growth, is well-defined even if beginning period imports are zero (so that ordinary growth would be infinite). It ranges from a minimum of -200 when end of period imports are zero to +200 when beginning period imports are zero. It is related to the usual measure of growth *g* by

$$G = \frac{200g}{200+g}$$

which, for moderate values of *g*, means that *G* is almost the same as *g*. As seen in Table 2, bounded growth is far less skewed than ordinary growth, with a mean of 25% and a median of 40% (note that median for bounded growth is virtually the same as the median for ordinary growth).

Tables 3 through 5 are the core of our data analysis. We focus on the estimates for bounded growth, but report results for ordinary growth for completeness. The top panel of Table 3 reports some descriptive regressions that illustrate the correlations in the complete sample as well as the non-quota-constrained sample. Our central specification is reported in the second panel of Table 3, where we regress bounded growth on the proximity-replenishment interaction and the change in trade barriers, including a complete set of country and product fixed effects (this is the specification given in equation 17b). We calculate *t*-statistics three ways: the usual OLS formula, White heteroskedasticity-consistent (labeled "robust std. errs."), and bootstrap. We also report results from a robust regression estimator, which is an iterative weighted least squares procedure that endogenously downweights outliers. The inference is the same across these four estimators: the proximity-replenishment effect β is about one, with a *t*-statistic above 3. How big is this effect? Since the range of the replenishment variable is between 0 and 67 percent, an estimated β of 1.04 implies that high-replenishment products from nearby countries grew 1.04×67 = 70 percentage points faster than otherwise. This is a big effect: it is more than 2.5 times faster than the mean level of bounded growth, and almost half again as fast as median growth. For products where replenishment is less important, with a replenishment percentage of 25%, the estimates still imply a big proximity effect, with imports growing 26 percentage points faster from nearby countries than more remote sources.

A feature of our data analysis is that we are pooling across trade flows of very different sizes, so it is of interest to look at the sensitivity of our results to regression weighting. The final two rows of Table 3 report weighted least squares results. When weighted by beginning-period imports, the inference about β is not much changed, with a point estimate of 0.85. Weighting by end of period imports, however, reverses the inference, indicating a small negative effect.

It is instructive to compare the replenishment-proximity effect with the effect of falling trade frictions. Multiplying the estimated trade friction semi-elasticity of -1.47 by the mean drop in trade frictions of 4.5 percentage points gives an effect on growth of 5.7 percentage points: a substantial effect, but small relative to mean import growth and the size of the replenishment-proximity effect.

Table 4 estimates the same specifications as Table 3, except ordinary rather than bounded growth is the dependent variable. The OLS estimate for β is an outlandish 787, although its standard error is quite large. Taking this estimate at face value implies an absurd 53,000 percentage point effect of proximity on import growth of high-replenishment products. Scaling this effect by the standard deviation of import growth (from Table 2) makes this number somewhat more meaningful, and implies that imports of high-replenishment goods from nearby-countries grew 0.64 standard deviations faster than from remote sources. The robust regression estimator, which effectively throws out extremely large and small values of growth because they are such outliers from the OLS line, delivers a result quite comparable to the results from Table 4, with an estimated β of 1.50.

The instability of the results of Table 4 induces a suspicion that a small number of outliers are driving the results, and we check this in Table 5. We identify outliers from first-stage regressions using the *DFITS* statistic, discard values for which $DFITS > 2\sqrt{\frac{k}{N}}$, and re-estimate the equations (see Belsley, Kuh, and Welsch, 1980, for the logic behind this procedure). The top panel of Table 5 shows that this procedure identifies a number of outliers when the dependent variable is bounded growth, but that inferences about the size of β are hardly affected. The bottom panel, by contrast, shows that inferences when the dependent variable is regular growth are completely dominated by a tiny number of outliers: dropping just 22 observations (0.8% of the sample) makes the estimated β statistically insignificant. Robust regression, which iteratively weights the remaining observations, yields a plausible point estimate of β=1.6, close to the estimates for β from the bounded growth regressions.

Our conclusion from the data analysis is that β is close to one. This is a big effect, both relative to the variation in the data and relative to the effect of falling trade frictions. While our empirical model cannot shed light on other determinants of changing trade flows, such as shifting comparative advantage and changes in quotas, it does control for them statistically. Overall, our results are consistent with the theoretical model: an increased demand for timeliness by retailers has led to a noticeable shift in trade patterns, with rapid-replenishment goods increasingly sourced from nearby countries.

5 Summary and conclusions

This paper has discussed some general equilibrium implications of the aphorism "time is money". In our model the demand for timeliness arises from variability in final demand, and we have showed that this has implications for international specialization: countries that are located close to major markets will have higher wages because they specialize in getting goods to market quickly. That countries close to the core are better off than peripheral countries is a common implication of economic geography models both new and old (see Fujita et al for an inventory), but our mechanism is new, and does not rely on the usual assumptions of transport costs and increasing returns to scale. Our model also offers an alternative explanation for the powerful effect of distance in empirical gravity equations: distance is proxying for time to market, not shipping costs.

We looked at data on the evolution of apparel imports into the US to see if an increased demand for timeliness has affected the pattern of trade. The answer is yes: imports of products where timeliness is important grew much faster from nearby countries than they did from the traditional sources of US apparel imports in East and South Asia.

The core idea behind the paper is that time matters, a cliche that has major implications for economic geography but which seems to have been neglected by theorists. We have developed some implications for international trade and inequality, but the general idea can be used to model agglomeration and regional inequality as well. It should also prove useful to develop models where the demand for timeliness comes from producers rather than final consumers, as suggested by the increasing importance of "just-in-time" inventory management practices.

Figure 1

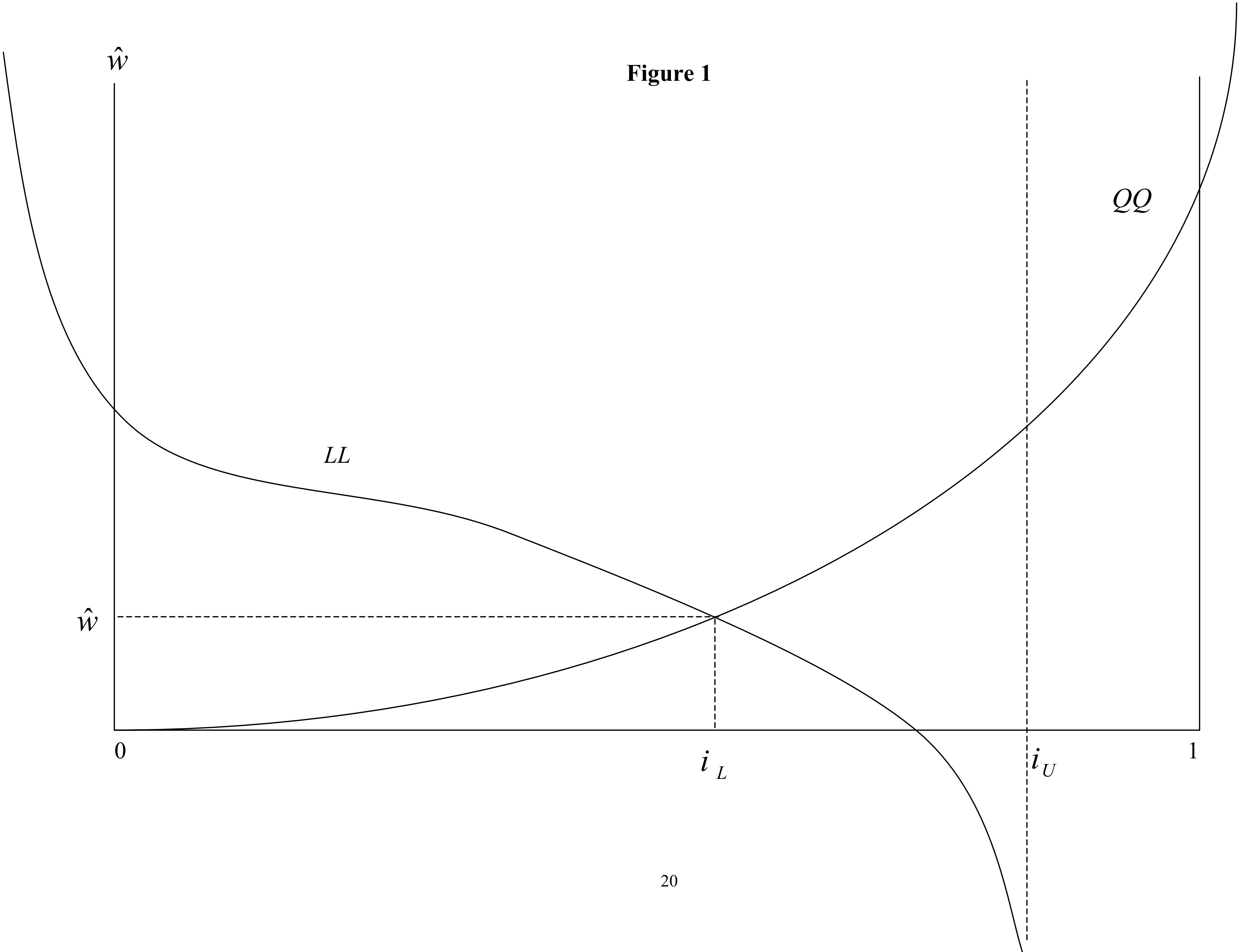
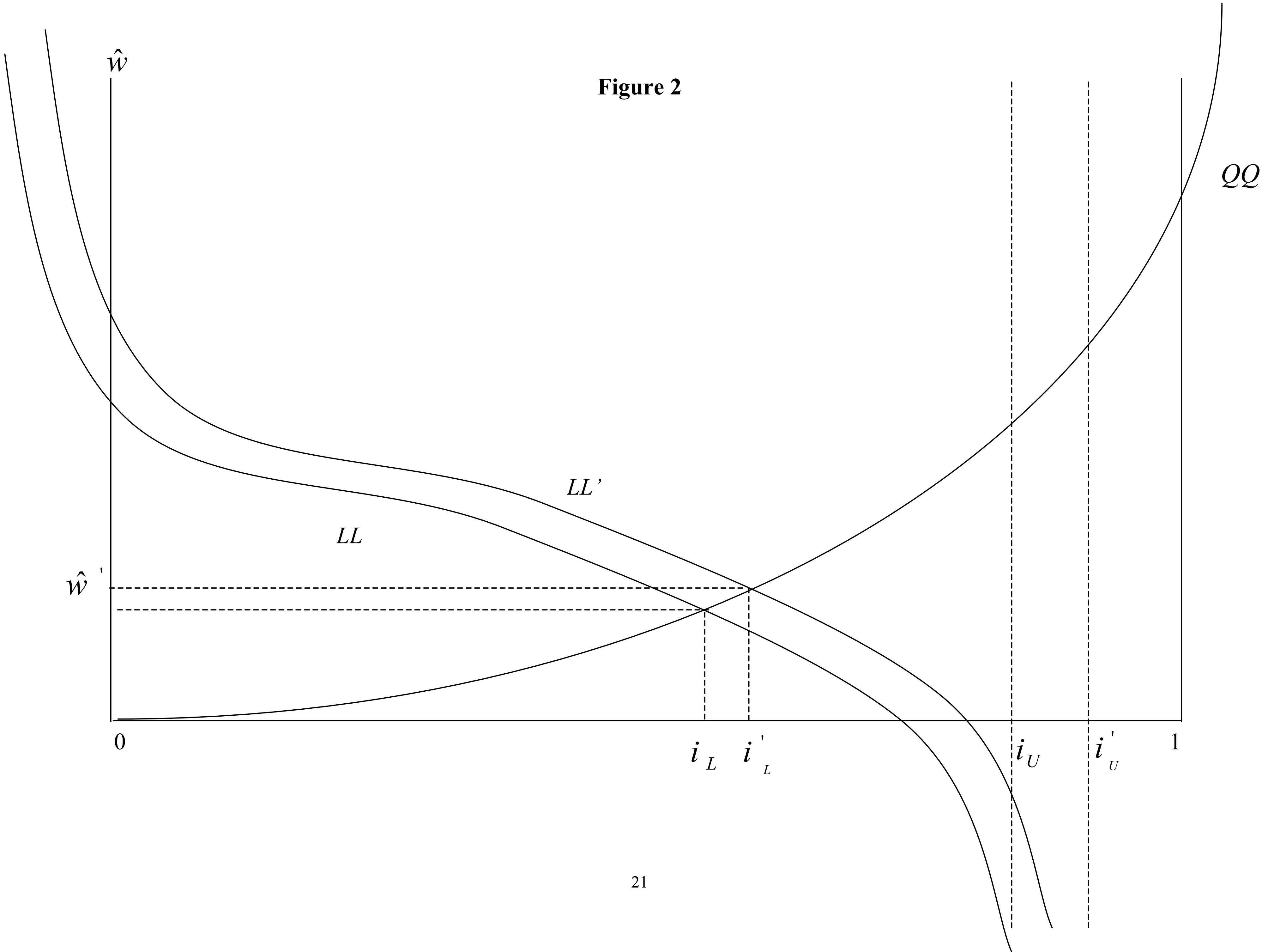


Figure 2



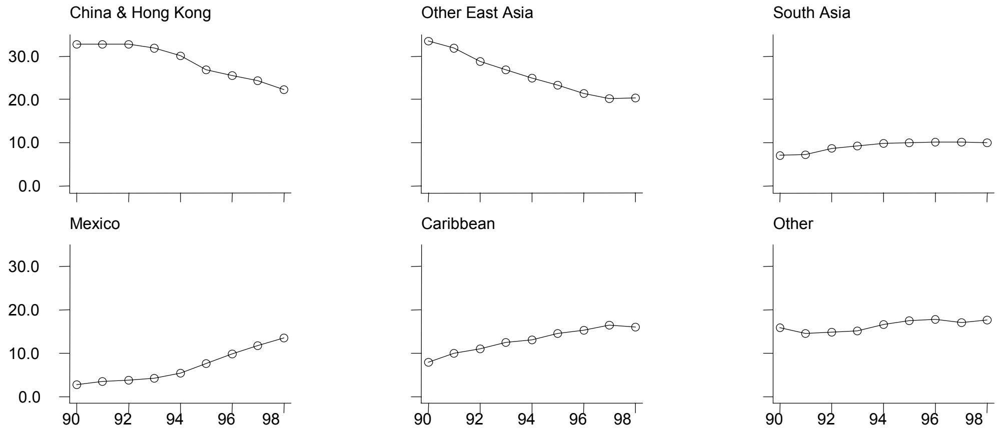


Figure 3 - Market Shares for US Apparel Imports, 1990-98

Table 1

Relative manufacturing wages, Mexico/China

1991	8.76
1992	
1993	6.72
1994	
1995	3.69
1996	3.27
1997	3.50
1998	3.20

Source: International Labour Organisation

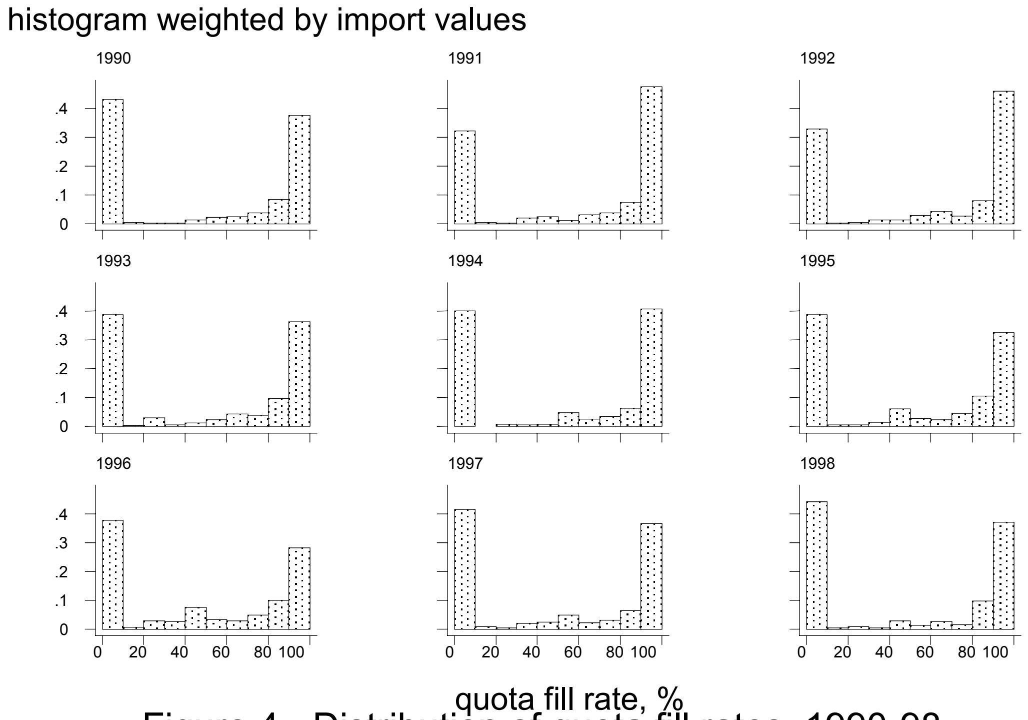


Figure 4 - Distribution of quota fill rates, 1990-98

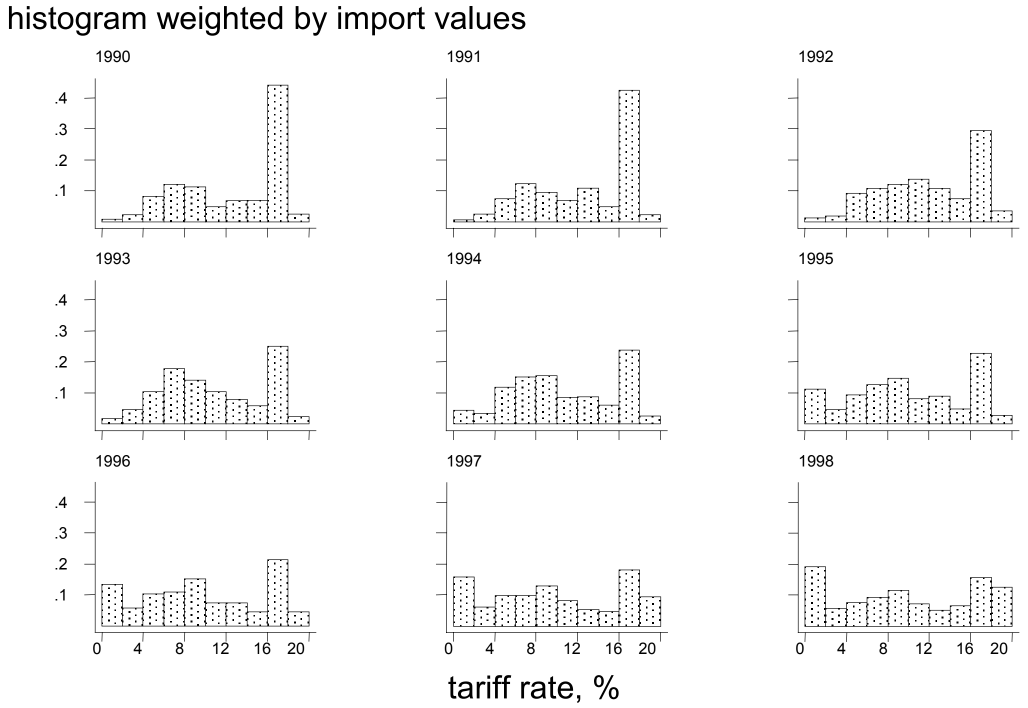


Figure 5 - Distribution of tariff rates, 1990-98

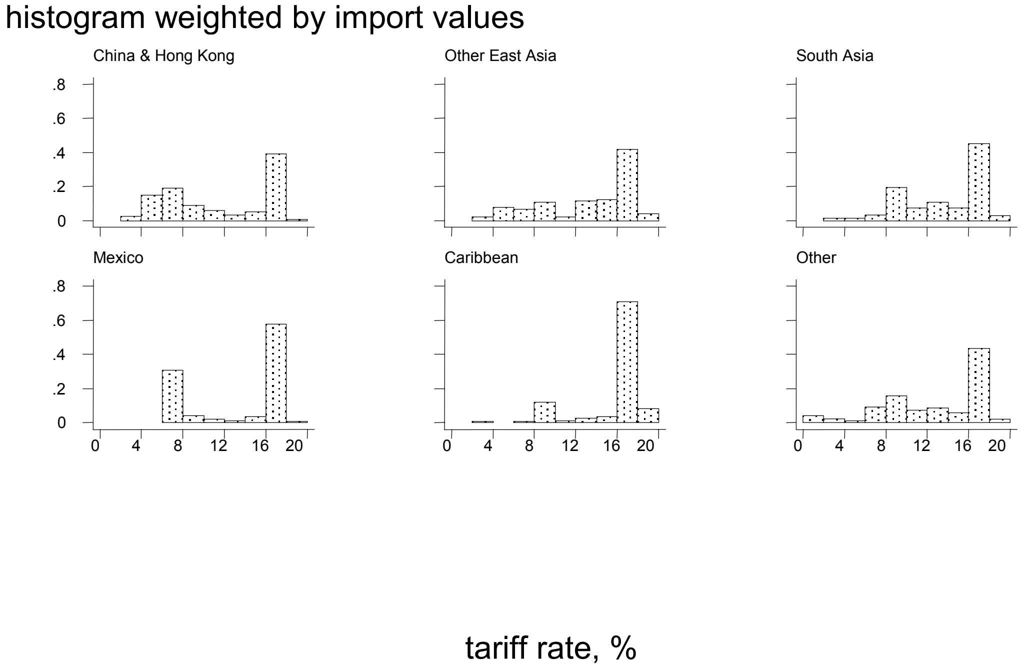


Figure 6 - Tariff incidence by region, 1990

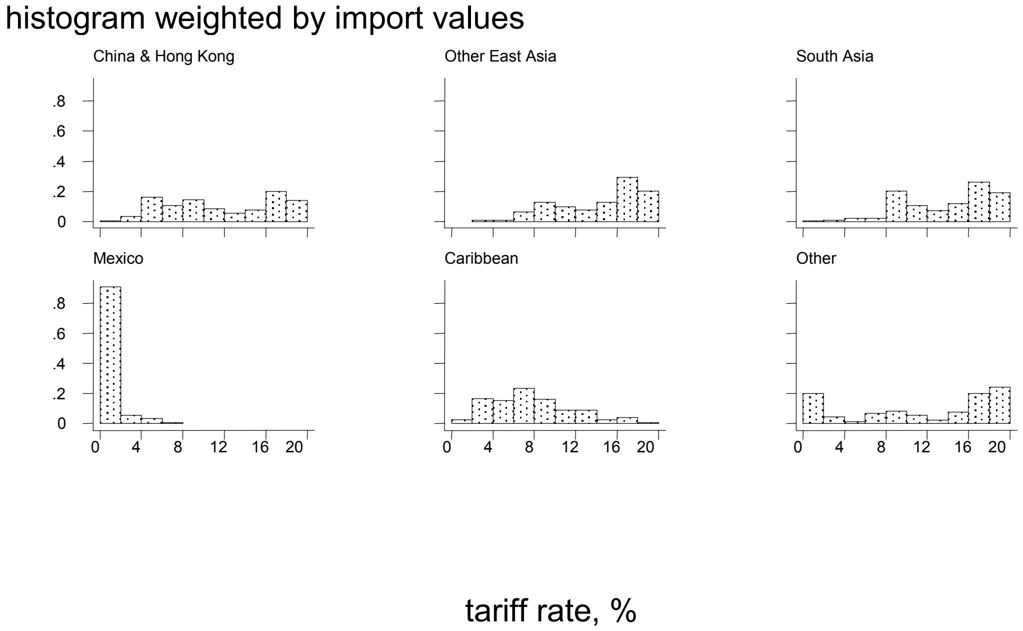


Figure 7 - Tariff incidence by region, 1998

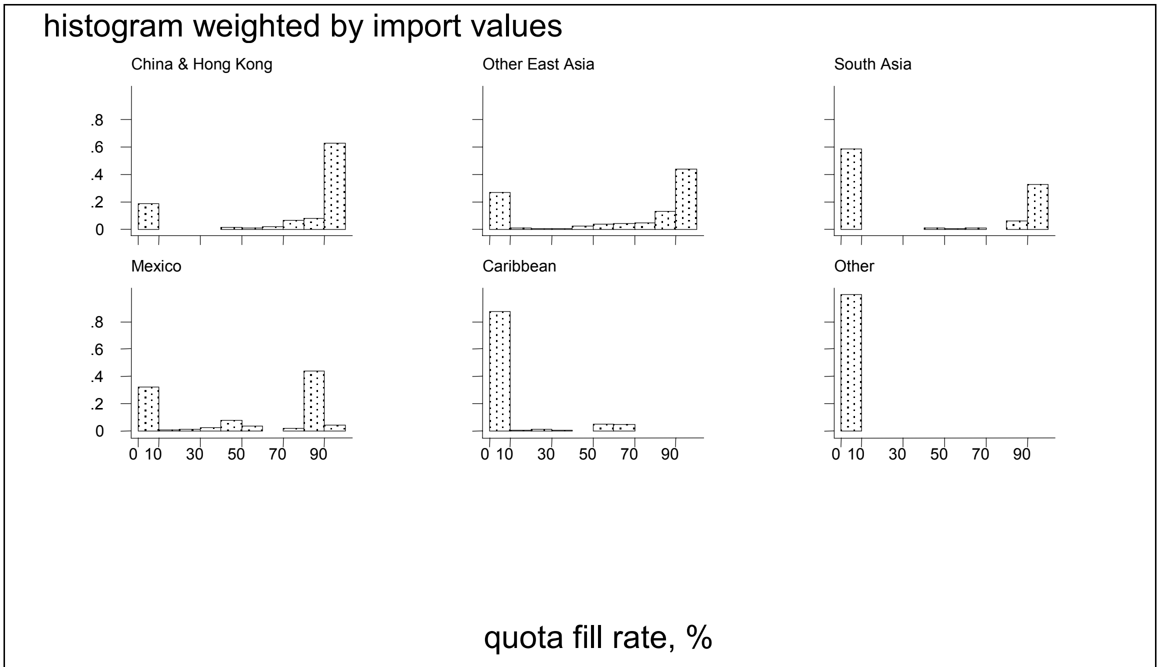


Figure 8 - Quota incidence by region, 1990

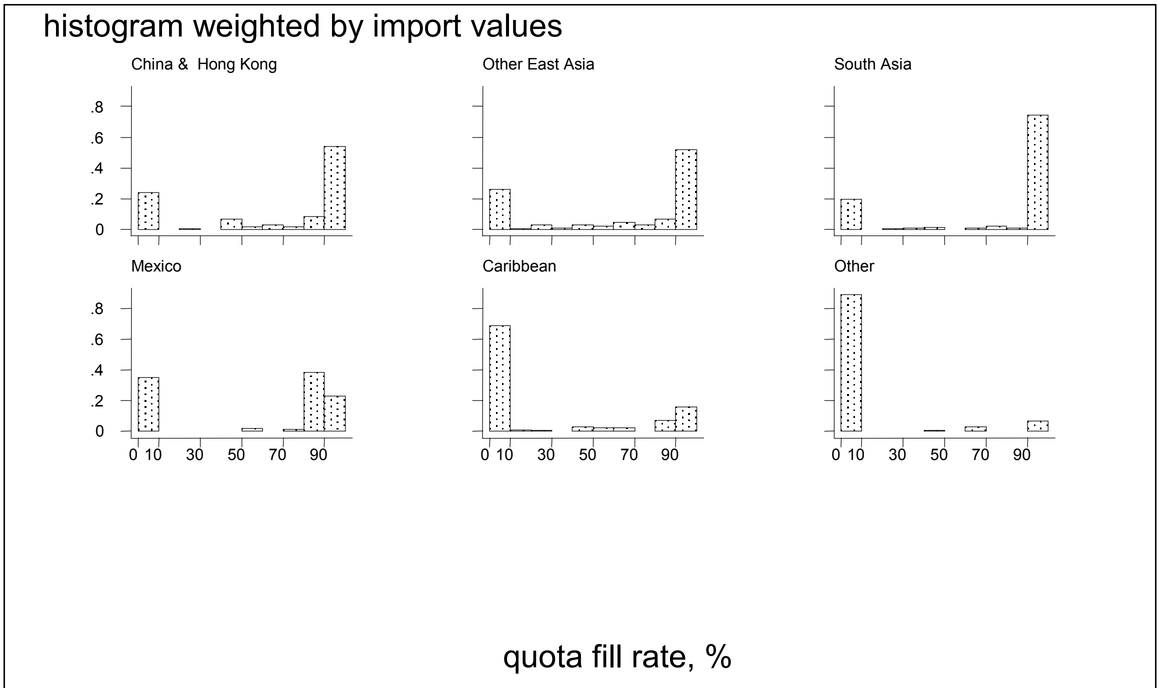


Figure 9 - Quota incidence by region, 1998

Table 2 - Descriptive Statistics

	Mean	Median	Std. Dev.	Min	Max
full sample (N = 3,177)					
import growth, %	6,395	49.7	82,123	-100	3,657,900
bounded import growth, %	25.8	40.0	123	-200	200
change in trade barriers, % points	-4.52	-3.00	14.2	-136	230
change in tariffs, % points	-2.77	-0.89	7.8	-83.7	33.3
change in transport costs, % points	-1.75	-1.22	11.4	-128	238
not quota constrained sample (N= 2,753)					
import growth, %	6,763	50	84,407	-99	3,657,900
bounded import growth, %	24.6	40.3	127	-200	200
change in trade barriers, % points	-4.54	-2.9	15	-136	230
change in tariffs, % points	-2.94	-0.9	8.2	-83.7	33.3
change in transport costs, % points	-1.61	-1.2	11.4	-128	238

Table 3 - Bounded import growth 1991-1998

3a exploratory OLS regressions, *t*-statistics in italics.

sample	Fixed effects?	proximity* replenishment	trade barriers
all (N = 3,177)	none	0.854 *3.40*	-1.726 *-11.5*
	country	0.920 *3.17*	-1.444 *-8.99*
	product	0.877 *3.34*	-1.654 *-11.1*
	country, product	0.982 *3.18*	-1.292 *-8.21*
not quota constrained only (N=2,753)	none	0.869 *3.25*	-1.676 *-10.6*
	country	0.965 *3.16*	-1.414 *-8.41*
	product	0.873 *3.09*	-1.608 *-10.3*

3b Central specification - all regressions include country and product fixed effects. Sample is observations not constrained by quotas (N = 2,753). *t*-statistics in italics.

estimator	proximity* replenishment	trade barriers
OLS, classical std. errs.	1.044 *3.15*	-1.269 *-7.66*
OLS, robust std. errs.	1.044 *3.75*	-1.269 *-7.08*
OLS, bootstrap std. errs.	1.044 *3.79*	-1.269 *-6.81*
Robust regression	1.051 *3.01*	-1.362 *-7.81*
weighted by 1991 imports	0.844 *3.90*	-2.061 *-5.70*
weighted by 1998 imports	-.375 *-2.77*	-1.135 *-5.68*

Notes to Table N: Dependent variable is bounded import growth between 1991 and 1998:

$$G_{ic} = 200 \cdot \frac{m_{ict} - m_{ic,t-1}}{\left(m_{ict} + m_{ic,t-1}\right)}$$

Table 4 - Import growth 1991-1998

4a Exploratory OLS regressions, *t*-statistics in italics.

sample	Fixed effects?	proximity* replenishment	trade barriers
all (N = 3,172)	none	482.8	-288.4
		2.81	*-3.23*
	country	712.2	-457.4
		3.23	*-3.74*
	product	511.3	-391.8
		2.77	*-3.74*
	country, product	738.2	-462.8
		3.02	*-3.71*
not quota constrained only (N=2,748)	none	514.3	-389.9
		2.83	*-3.64*
	country	750.2	-488.0
		3.29	*-3.88*
	product	541.7	-401.3
		2.76	*-3.69*

4b Central specification - all regressions include country and product fixed effects. Sample is observations not constrained by quotas (N = 2,748). *t*-statistics in italics.

estimator	proximity* replenishment	trade barriers
OLS, classical std. errs.	787.0	-491.7
	3.06	*-3.83*
OLS, robust std. errs.	787.0	-491.7
	1.72	*-1.97*
OLS, bootstrap std. errs.	787.0	-491.7
	1.80	*-1.99*
Robust regression	1.5	-1.4
	2.92	*-5.49*
weighted by 1991 imports	3.0	-13.2
	0.44	*-1.15*
weighted by 1998 imports	15,080	-7470
	5.40	-18.1

Notes to Table 4: Dependent variable is percentage import growth between 1991 and 1998:

$$g_{ic} = 100\frac{m_{ict} - m_{ic,t-1}}{m_{ic,t-1}}$$

Table 5 - Sensitivity to outliers

5a Bounded import growth, central specification without outliers - all regressions include country and product fixed effects. 82 outliers deleted (N = 2,671). *t*-statistics in italics.

estimator	proximity* replenishment	trade barriers
OLS, classical std. errs.	1.134 *3.58*	-1.468 *-8.63*
OLS, robust std. errs.	1.134 *4.20*	-1.468 *-7.80*
OLS, bootstrap std. errs.	1.134 *4.07*	-1.468 *-7.25*
Robust regression	1.097 *3.27*	-1.526 *-8.50*
weighted by 1991 imports	0.805 *3.71*	-2.116 *-5.81*
weighted by 1998 imports	-0.429 *-3.16*	-1.117 *-5.73*

5b regular import growth, central specification without outliers - all regressions include country and product fixed effects. 22 outliers deleted (N = 2,726). *t*-statistics in italics.

estimator	proximity* replenishment	trade barriers
OLS, classical std. errs.	24.1 *0.62*	-109 *-5.38*
OLS, robust std. errs.	24.1 *0.47*	-109 *-3.01*
OLS, bootstrap std. errs.	24.1 *0.48*	-109 *-3.02*
Robust regression	1.6 *3.08*	-1.7 *-6.35*
weighted by 1991 imports	-1.1 *-0.47*	-11.2 *-2.99*
weighted by 1998 imports	-5.0 *-0.20*	-230 *-5.26*

Notes to Table 5: Outliers are unusually influential observations, as defined by the *DFFITS* statistic computed in the central specification. See the text for details.

References

Abernathy, Frederick H., John T. Dunlop, Janice H. Hammond, and David Weil, *A Stitch in Time: Lean Retailing and the Transformation of manufacturing - Lessons from the apparel and textile industries*, 1999, Oxford and New York: Oxford University Press.

Belsley, D.A., E. Kuh, and R. E. Welsch, 1980, *Regression Diagnostics*, New York: John Wiley.

Deardorff, Alan V., 2002, "Time and Trade: The Role of Time in Determining the Structure and Effects of International Trade, with an Application to Japan," in Robert M. Stern, ed., *Analytical Studies in U.S.-Japan International Economic Relations*, Cheltenham, U.K. and Northhampton, MA: Edward Elgar Publishing Inc.

Deardorff, Alan V., 2001, “Local comparative advantage: trade costs and the pattern of trade”, manuscript, University of Michigan.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables, 1999, *The Spatial Economy: Cities, Regions, and International Trade*, Cambridge, MA: MIT Press.

Davis, Steven, Haltiwanger, John R., and Scott Schuh, 1996, *Job Creation and Destruction*, Cambridge, MA: MIT Press.

Hummels, David, "Time as a trade barrier", 2001, manuscript, Purdue University.

Redding, Steven R., and Anthony J. Venables, "Economic geography and international inequality", 2001, manuscript, LSE.

Venables, Anthony J., 2001, "Geography and international inequalities: the impact of new technologies", *Journal of Industry, Competition and Trade*, 1, 135-160 (June).

Appendix

This mathematical appendix works out the details of the analysis in the text, including some parameter restrictions required for the model to make sense. It has four sections:

A1 Optimal production plans for risk-neutral firms.

A2 Locational equilibrium, including the properties of the *QQ* curve.

A3 Labor market equilibrium, including the properties of the *LL* curve.

A4 Comparative statics of location and wages in general equilibrium.

A1 The firm's problem

As discussed in the text, we consider risk-neutral firms who face two consecutive realizations of demand, and must decide how much to produce and how much to sell in each period. The timing is as follows:

1) All firms decide how much to produce
2) Demand level for period 1 is realized, and firms decide how much to sell. Any output not sold can be held until period 2.
3) Flexible firms produce again; nonflexible firms do not
4) Demand level for period 2 is realized, and firms decide how much to sell. Any output not sold is thrown out.

We consider the nonflexible firm's problem first. We begin with the problem facing the firm after it has already produced, and then work out optimal production given the solution. After producing some level of output and observing a_1, output costs are sunk, so the objective is simply to maximize expected revenue, subject to the constraint that total sales not exceed output. Both revenue and the shadow value of second period output depend on the realization of a_2. As a result, the constrained maximization problem is to choose $\{s_1, s_2^H, s_2^L\}$ to maximize

$$L = s_1(a_1 - bs_1) + \rho s_2^H(a_H - bs_2^H) + (1-\rho)s_2^L(a_L - bs_2^L) + \lambda_H(q - s_1 - s_2^H) + \lambda_L(q - s_1 - s_2^L)$$

where ρ is the probability that $a_i = a_H$. This is a complementary slackness problem, for which the first order conditions and associated solutions are

$$\frac{\partial L}{\partial s_1} = a_1 - 2bs_1 - \lambda_H - \lambda_L = 0 \quad \rightarrow \quad s_1 = \frac{a_1}{2b} - \frac{(\lambda_H + \lambda_L)}{2b}$$

$$\frac{\partial L}{\partial s_2^H} = \rho a_H - 2\rho b s_2^H - \lambda_H = 0 \quad \rightarrow \quad s_2^H = \frac{a_H}{2b} - \frac{\lambda_H}{2\rho b}$$

$$\frac{\partial L}{\partial s_2^L} = (1-\rho) a_L - 2(1-\rho) b s_2^L - \lambda_L = 0 \quad \rightarrow \quad s_2^L = \frac{a_L}{2b} - \frac{\lambda_L}{2(1-\rho) b}$$

There are four possible configurations for the Lagrange multipliers:

A. $\lambda_H > 0, \lambda_L > 0$

B. $\lambda_H > 0, \lambda_L = 0$

C. $\lambda_H = 0, \lambda_L > 0$

D. $\lambda_H = 0, \lambda_L = 0$.

The logic of the model means we can dismiss cases C and D. Case A is the simplest, and corresponds to the solutions in the text, which are

$$s_1 = \frac{q}{2} + \frac{a_1 - \bar{a}}{4b}, \tag{A1}$$

$$s_2 = \frac{q}{2} - \frac{a_1 - \bar{a}}{4b}. \tag{A2}$$

We can also calculate the value of the Lagrange multipliers in case A,

$$\lambda^H = \rho\left[\frac{2a_H + a_1 - \bar{a}}{2} - bq\right], \tag{A3}$$

$$\lambda^L = (1-\rho)\left[\frac{2a_L + a_1 - \bar{a}}{2} - bq\right] \tag{A4}$$

In case B, setting $\lambda_L = 0$ immediately implies

$$s_2^L = \frac{a_L}{2b} \tag{A5}$$

Solving for the other Lagrange multiplier gives

$$\lambda_H = \rho\left[a_H - 2b(q - s_1)\right] \tag{A6}$$

Plugging this into the expressions for first period sales and second period high-demand sales gives

$$s_1 = \frac{1}{1+\rho}\left[\rho q + \frac{a_1 - \rho a_H}{2b}\right] \tag{A7}$$

$$s_2^H = \frac{1}{1+\rho}\left[q - \frac{a_1 - \rho a_H}{2b}\right] \tag{A8}$$

To interpret these expressions, consider first the case where $\rho = 0$ for period 2. This means that the high state never happens, so marginal revenue is always zero in period 2. As a result, first-period sales are $\frac{a_1}{2b}$, the level of sales that sets marginal revenue to zero in period 1. Looking at the case where $\rho = 1$ for period 2, we get

$$s_1 = \frac{q}{2} + \frac{a_1 - a_H}{4b}, \quad s_2^H = \frac{q}{2} - \frac{a_1 - a_H}{4b}$$

which implies that expected sales, and therefore expected marginal revenue, are equalized. If $a_1 = a_H$, output is just split equally; if $a_1 = a_L$, then some output is saved for the high demand period to come.

We can now solve for optimal output q, which maximizes expected revenue minus actual costs. In case A, where second period marginal revenue is always positive, substitution of (A1) and (A2) into the definition of revenue gives, after a bit of manipulation,

$$Expected\ revenue_A = const + \overline{a}q - \frac{bq^2}{2}$$

Plugging this into the defintion of profit and maximizing immediately yields the result

$$q^N = \frac{\overline{a} - w}{b} \tag{A9}$$

We can substitute this into the solutions (A3) and (A4) for the Lagrange multipliers to find their values at the optimum:

$$\lambda^H = \rho\left[\frac{2a_H + a_1 - 3\overline{a} + 2w}{2}\right] \tag{A10}$$

$$\lambda^L = (1-\rho)\left[\frac{2a_L + a_1 - 3\overline{a} + 2w}{2}\right] \tag{A11}$$

We can use these expressions to see for what range of parameters case A is relevant. A sufficient condition is that λ^L is always positive, which occurs if

$$w > \frac{3}{2}(\overline{a} - a_L) \tag{A12}$$

This can be guaranteed by a suitable parameter restriction. In particular, recall that in general equilibrium the price of the numeraire good is one, which places a lower bound on the nominal wage in general equilibrium. We can choose units of the numeraire good so that this lower bound is given by the right-hand side of the inequality (A12), and to keep things simple we do so.

Plugging (A9) into (A1) and (A2) gives the value of sales at the optimum:

$$s_1^N = \frac{\overline{a}-w}{2b} + \frac{a_1 - \overline{a}}{4b} = \frac{\overline{a}+a_1-2w}{4b} \tag{A13}$$

$$s_2^N = \frac{\overline{a}-w}{2b} - \frac{a_1 - \overline{a}}{4b} = \frac{3\overline{a}-a_1-2w}{4b} \tag{A14}$$

We know that the revenue-maximizing value of sales in period 2 when demand is low is $\frac{a_L}{2b}$. For (A14) to make sense, s_2^N must be less than this upper bound when first period demand is low, or

$$\frac{3\overline{a}-a_L-2w}{4b} \le \frac{a_L}{2b} \tag{A15}$$

The implied restriction on *w* is exactly the lower bound on *w* derived in (A12).

Moving to the solution for flexible firms, again we work backward. After first period demand is realized but before second period demand is realized, the firm chooses first-period sales and second-period output output and conditional sales to maximize expected profit subject to the constraint that sales are no greater than output. The maximand is

$$\begin{aligned} L =& s_1\left(a_1 - bs_1\right) + \rho s_2^H(a_H - bs_2^H) + \left(1-\rho\right)s_2^L(a_L - bs_2^L) - wq_2 \\ &+ \lambda_1(q_1 - s_1) + \lambda_H(q_1 + q_2 - s_1 - s_2^H) + \lambda_L(q_1 + q_2 - s_1 - s_2^L) \end{aligned} \tag{A16}$$

As for the nonflexible firm, this is a complementary slackness problem, for which the first order conditions and associated solutions are

$$\frac{\partial L}{\partial s_1} = a_1 - 2bs_1 - \lambda_H - \lambda_L - \lambda_1 = 0 \quad \rightarrow \quad s_1 = \frac{a_1}{2b} - \frac{\left(\lambda_H + \lambda_L + \lambda_1\right)}{2b}$$

$$\frac{\partial L}{\partial s_2^H} = \rho a_H - 2\rho b s_2^H - \lambda_H = 0 \quad \rightarrow \quad s_2^H = \frac{a_H}{2b} - \frac{\lambda_H}{2\rho b}$$

$$\frac{\partial L}{\partial s_2^L} = (1-\rho)a_L - 2(1-\rho)bs_2^L - \lambda_L = 0 \quad \rightarrow \quad s_2^L = \frac{a_L}{2b} - \frac{\lambda_L}{2(1-\rho)b}$$

$$\frac{\partial L}{\partial q_2} = -w + \lambda_H + \lambda_L = 0 \quad \rightarrow \quad \lambda_H + \lambda_L = w$$

For the moment, we assume that the first period output constraint is always slack, so that λ_1 =0. As before, there are four possible configurations for the Lagrange multipliers λ_H and λ_L. When both are positive the solution is

$$s_1^F = \frac{a_1 - w}{2b} \tag{A17}$$

$$s_2^F = \frac{\overline{a} - w}{2b} \tag{A18}$$

$$q_2^F = \frac{\overline{a} - w}{2b} - (q_1 - s_1) \geq 0 \tag{A19}$$

The Lagrange multipliers at the optimum depend on *w*, and will always be positive as long as inequality (A12) is satisfied. Since we have already assumed this, there is no need to discuss the other configurations of the Lagrange multipliers.

We now come to the last element of the problem, the choice of q_1. Expected first period revenue is highest if the multiplier λ_1 in (16) is always zero (as tentatively assumed), and since unsold first-period output can be sold in the second period, total profits will be maximized if λ_1 is always zero. This can be accomplished by producing at least enough in period 1 to sell (A17) if demand is high, which is

$$q_1^F = \frac{a_H - w}{2b} \tag{A20}$$

Another consideration is given by the constraint in (A19) that second period-output can't be negative. To check if (A19) is satisfied when first period demand is low, substitute (A20) and (A17) into (A19), setting $a_1 = a_L$. The result is

$$\overline{a} - w \geq a_H - a_L \tag{A21}$$

This inequality states that average demand must be sufficiently large relative to the variance in demand, which we assume to hold in equilibrium.

Finally, we note that there is some indeterminacy in the solution for optimal first period output. Equation (A20) gives the minimum output level to guarantee that first-period sales are ex-post optimal, but since sales continue in the second period, the firm may choose to produce more in period 1 and less in period 2. To resolve this indeterminacy, we assume very small storage costs, which will lead the firm to produce q_1^F without affecting anything else about the problem.

A2 Locational equilibrium

The choice of production location involves a tradeoff between the benefits of flexibility and higher wage costs. To derive the cutoff point, we calculate expected profits in each location as a function of the variance of output, with outputs chosen optimally as outlined above. By assumption, producers in in *C* are flexible, while producers in *A* are not.

Expected profits for producers in *A* are expected revenues minus costs. Costs are found by multiplying wages in *A* by optimal output as given by (4):

$$costs_A = \frac{\overline{a}w_A - w_A^2}{b} \qquad \text{(A22)}$$

Expected revenue is computed by substituting optimal sales choices into the definition of revenue, and using the fact that

$$V(a) = \rho a_H^2 + (1-\rho)a_L^2 - \overline{a}^2 . \qquad \text{(A23)}$$

The result is that expected revenue in *A* is

$$revenue_A = \frac{4\overline{a}^2 - 4w_A^2 + V(a)}{8b} . \qquad \text{(A24)}$$

Similarly, expected costs and revenue in *C* are calculated as

$$costs_C = \frac{\overline{a}w_C - w_C^2}{b} \qquad \text{(A25)}$$

$$revenue_C = \frac{4\overline{a}^2 - 4w_C^2 + 2V(a)}{8b} \qquad \text{(A26)}$$

In both (A24) and (A26), revenue is increasing in the variance of demand, which is a consequence of the convexity of the profit function. Both types of producers respond to period 1 shocks, but producers in *C* respond more, so expected revenues in *C* increase faster as a function of *V*(*a*) than they do in *A*. It is clear by inspection that if wages are equal in the two locations then profits will be higher in *C*, and it is also clear that, holding *V*(*a*) constant, a big enough wage premium in *C* will cause profits there to fall below profits in *A*. Finally, for a given wage premium in *C*, there is a level of *V*(*a*) which will equalize profits.

To find the critical value of *V*(*a*), we substitute equation (7) into (A24) and (A26), substitute (A22) and (A24)-(A26) into the definition of profits, and set profits in the two locations equal. Solving for *i* gives the result for i_L in the text, equation (8). The total derivative of i_L is

$$di_L = \frac{8}{\sigma^2}\left[\left(\overline{a}-w_C\right)dw_C - \left(\overline{a}-w_A\right)dw_A\right] \tag{A27}$$

and the partial derivatives are

$$\frac{\partial i_L}{\partial w_C} = \frac{8\left(\overline{a}-w_C\right)}{\sigma^2} \tag{A28}$$

$$\frac{\partial i_L}{\partial w_A} = -\frac{8\left(\overline{a}-w_A\right)}{\sigma^2} \tag{A29}$$

As long as $\overline{a} > w_C$ and $\overline{a} > w_A$, which we have already assumed (and which simply means that demand is high enough for the model to make sense), we get the expected result that i_L is increasing in w_C and decreasing in w_A. But it is not possible to solve for i_L as a function of the wage differential $\hat{w} = w_C - w_A$, since (8) is a quadratic in w_C and w_A separately. With the restrictions that the wage differential cannot be negative in equilibrium and that $i_L \in [0,1]$, we can depict equation (8) as a surface in i_L- w_C-w_A space, as in figure A1. The surface is a quadratic, increasing at a decreasing rate in w_C and decreasing at a decreasing rate in w_A. The *QQ* curve of Figures 1 and 2 is traced out by the equilibrium movements of w_C and w_A as i_U changes, which are derived in the next section of this appendix. In general all we know is that the relationship is monotonically increasing in w_C-w_A along the equilibrium path, so the shape of *QQ* in Figures 1 and 2 is otherwise free hand.

A3 Labor Market equilibrium

For given i_U and i_L, there are i_U - i_L firms that locate in *C*, with the remaining 1 - (i_U - i_L) firms located in *A*. Each firm in each location has average annual labor demand of $2q^*$, where

$$q^* = \frac{\overline{a}-w}{2b}, \tag{A30}$$

so total labor demand in *C* and *A* respectively is

$$(i_U - i_L)\frac{\overline{a} - w_C}{b} \tag{A31}$$

$$(1 + i_L - i_U)\frac{\overline{a} - w_A}{b} \tag{A32}$$

Setting labor supply equal to labor demand in each region and solving for wages gives equations (9) and (10) in the text, and subtracting (11) from (9) gives (12). The vertical intercept of (12) at $i_L = 0$ is

$$\hat{w} = b\left(\frac{L_A}{1 - i_U} - \frac{L_C}{i_U}\right), \tag{A33}$$

which is positive due to the parameter restriction given by equation (10). The economically relevant range of the function is $[0, i_U)$, and $\hat{w}$ asymptotically approaches minus infinity as $i_L \rightarrow i_U$ from below.

From equation (12), the slope of $\hat{w}$ when graphed against i_L is

$$\frac{\partial \hat{w}}{\partial i_L} = -b\left(\frac{L_A}{(1 + i_L - i_U)^2} + \frac{L_C}{(i_U - i_L)^2}\right) \tag{A34}$$

which is strictly negative. The second derivative is

$$\frac{\partial^2 \hat{w}}{\partial i_L^2} = 2b\left(\frac{L_A}{(1 + i_L - i_U)^3} - \frac{L_C}{(i_U - i_L)^3}\right) \tag{A35}$$

which is first positive then negative. As a consequence, the *LL* curve has the shape shown in Figures 1 and 2.

A4 Comparative statics

Our core comparative static experiment is an increase in the range of products for which flexible production is feasible. We model this through an increase in i_U, which has two effects. The first is straightforward, which is a horizontal shift of Δi_U in the *LL* curve (to prove this, totally differentiate (12), set $\hat{w} = 0$, and solve for $\mathrm{d}i_L/\mathrm{d}i_U = 1$). The second part of the story is that an increase in i_U changes the *QQ* curve, which makes analysis somewhat tricky.

We start with the simplest case. Here we suppose that there is a one-to-one negative relationship between selling season and demand variance, so that (7) is valid for all $i \in [0,1]$. In this case an increase in i_U simply extends the upper boundary of the *QQ* relationship, without changing its shape, which is the case analyzed graphically in Figure 2. This means that we can use calculus to analyze the comparative statics. We have three equations (9), (11) and (8) in the three unknowns w_A, w_C, and i_L. Excluding parameters, the exogenous variables that can change the equilibrium are L_A, L_C, and i_U.

Totally differentiating the three equations gives

$$dw_C = \frac{-b}{i_U - i_L} dL_C + \frac{bL_C}{\left(i_U - i_L\right)^2} di_U - \frac{bL_C}{\left(i_U - i_L\right)^2} di_L \qquad (9')$$

$$dw_A = \frac{-b}{1 + i_L - i_U} dL_A - \frac{bL_A}{\left(1 + i_L - i_U\right)^2} di_U + \frac{bL_A}{\left(1 + i_L - i_U\right)^2} di_L \qquad (11')$$

$$di_L = \frac{16\left(\overline{a} - w_C\right)}{3\sigma^2} dw_C - \frac{16\left(\overline{a} - w_A\right)}{3\sigma^2} dw_A \qquad (8')$$

Defining $\theta = i_U - i_L$, $\beta_C \equiv \frac{16}{3\sigma^2}\left(\overline{a} - w_C\right)$, $\beta_A \equiv \frac{16}{3\sigma^2}\left(\overline{a} - w_A\right)$ and bringing the endogenous variables to the left hand side, gives

$$dw_C + \frac{bL_C}{\theta^2} di_L = \frac{-b}{\theta} dL_C + \frac{bL_C}{\theta^2} di_U \qquad (9')$$

$$dw_A - \frac{bL_A}{\left(1-\theta\right)^2} di_L = -\frac{b}{1-\theta} dL_A - \frac{bL_A}{\left(1-\theta\right)^2} di_U \qquad (11')$$

$$di_L - \beta_C dw_C + \beta_A dw_A = 0 \qquad (8')$$

Writing this system out in matrix notation,

$$\begin{bmatrix} 1 & 0 & \frac{bL_C}{\theta^2} \\ 0 & 1 & \frac{-bL_A}{\left(1-\theta\right)^2} \\ -\beta_C & \beta_A & 1 \end{bmatrix} \begin{bmatrix} dw_c \\ dw_A \\ di_L \end{bmatrix} = \begin{bmatrix} \frac{-b}{\theta} & 0 & \frac{bL_C}{\theta^2} \\ 0 & \frac{-b}{1-\theta} & \frac{-bL_A}{\left(1-\theta\right)^2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} dL_C \\ dL_A \\ di_U \end{bmatrix} \qquad (A36)$$

Using the shorthand $\mathbf{A}dy = \mathbf{B}dx$ to represent the system (A36) above, we have

$$\det(\mathbf{A}) = 1 + \frac{\beta_C bL_C}{\theta^2} + \frac{\beta_A bL_A}{\left(1-\theta\right)^2} > 0 \qquad (A37)$$

Solving the system (A36) for *dy*=$\mathbf{A}^{-1}\mathbf{B}dx$ gives

$$\begin{bmatrix} dw_C \\ dw_A \\ di_L \end{bmatrix} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} 1+\dfrac{\beta_A bL_A}{(1-\theta)^2} & \dfrac{\beta_A bL_c}{\theta^2} & \dfrac{-bL_c}{\theta^2} \\ \dfrac{\beta_C bL_A}{(1-\theta)^2} & 1+\dfrac{\beta_C bL_c}{\theta^2} & \dfrac{bL_A}{(1-\theta)^2} \\ \beta_C & -\beta_A & 1 \end{bmatrix} \begin{bmatrix} \dfrac{-b}{\theta} & 0 & \dfrac{bL_c}{\theta^2} \\ 0 & \dfrac{-b}{1-\theta} & \dfrac{-bL_A}{(1-\theta)^2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} dL_c \\ dL_A \\ di_U \end{bmatrix} \tag{A38}$$

or equivalently,

$$\begin{bmatrix} dw_C \\ dw_A \\ di_L \end{bmatrix} = \frac{b}{\det(\mathbf{A})} \begin{bmatrix} \dfrac{-1}{\theta}\left(1+\dfrac{\beta_A bL_A}{(1-\theta)^2}\right) & \dfrac{-\beta_A bL_C}{(1-\theta)\theta^2} & \dfrac{L_C}{\theta^2} \\ -\dfrac{\beta_C bL_A}{\theta(1-\theta)^2} & \dfrac{-1}{1-\theta}\left(1+\dfrac{\beta_C bL_c}{\theta^2}\right) & \dfrac{-L_A}{(1-\theta)^2} \\ \dfrac{-\beta_C}{\theta} & \dfrac{\beta_A}{1-\theta} & \dfrac{\beta_C L_c}{\theta^2}+\dfrac{\beta_A L_A}{(1-\theta)^2} \end{bmatrix} \begin{bmatrix} dL_c \\ dL_A \\ di_U \end{bmatrix} \tag{A39}$$

Scrutinizing the solution establishes

$$\begin{bmatrix} \dfrac{dw_C}{dL_C} & \dfrac{dw_C}{dL_A} & \dfrac{dw_C}{di_U} \\ \dfrac{dw_A}{dL_C} & \dfrac{dw_A}{dL_A} & \dfrac{dw_A}{di_U} \\ \dfrac{di_L}{dL_C} & \dfrac{di_L}{dL_A} & \dfrac{di_L}{di_U} \end{bmatrix} = \begin{bmatrix} - & - & + \\ - & - & - \\ - & + & + \end{bmatrix} \tag{A40}$$

These signs are all as expected. It is also the case that $di_L/di_U < 1$ (by inspection). All of this corresponds to the results from the graphical analysis.

The model is not susceptible to analysis using calculus when selling season and demand variance do not have a one-to-one relationship, but the substantive conclusions are the same. At any given relative wage $\hat{w}$ there is a mass of potentially flexible firms that wants to produce in *C*, and this mass is weakly decreasing in $\hat{w}$. This is because there is only so much labor in *C*, and the equilibrium $\hat{w}$ prices some of the potentially flexible firms out of the market, so that they produce in *A* instead. Now increase the mass of potentially flexible firms by the amount Δi_u. Of these new entrants into the market for *C* labor, some fraction γ will want to produce in *C* at the old equilibrium wage, since their revenue increase from flexible production exceeds the increased wage costs associated with *C*. This means there is excess demand for *C* labor at the old equilibrium $\hat{w}$, so w_C rises, choking off some of the increased demand for *C* labor. As a result, less than $\gamma\Delta i_u$ move to *C* on net. Since some firms have left the market for *A* labor, wages there must fall. Therefore, $\hat{w}$ rises. If $\gamma = 0$, there is no effect: all of the newly-flexible firms are content to stay in low-wage *A* even though flexible production is now feasible, since it is not profitable. If $\gamma = 1$, the algebraic results will be the same as found in equations (A39).

Figure A1 - QQ surface
(the origin is in the bottom rear corner of the box)