Quotas in General Equilibrium

David R. Baqaee
UCLA

Kunal Sangani
Stanford*

March 20, 2025

Abstract

We analyze economies with quotas and other quantity-based distortions. We show that any feasible, distorted allocation of resources can be implemented as the competitive equilibrium of an economy with quotas. In contrast to decentralization via wedges, economies with quotas are constrained efficient, and hence they satisfy macro-envelope conditions. This means that the effects of changes in distortions and policy reforms can be expressed in terms of a small set of sufficient statistics. We provide a non-parametric and nonlinear characterization of the effects of quota and productivity changes on aggregate output. We also calculate the costs of misallocation due to quotas. These costs are expressed in terms of the price elasticities of an inverse demand system capturing how quota prices respond to changes in quota levels. We illustrate our results using several examples: we estimate efficiency gains from increasing the cap on H-1B visas, relaxing zoning restrictions on single-family housing in American cities, removing capital controls in Argentina, phasing out U.S. quotas on Chinese clothing and textiles, and eliminating the taxicab medallion system in New York City. Our results offer a flexible method for quantifying the costs of quota distortions and the effects of policy reforms across these settings.

*Emails: baqaee@econ.ucla.edu, ksangani@stanford.edu. We thank Chang He, Pete Klenow, Pablo Fajgelbaum, and Oleg Itskhoki for valuable comments. We are also grateful to Judith Dean, Amit Khandelwal, and Peter Schott for sharing data on quota prices for Chinese clothing and textile exports.

1 Introduction

Quotas—and quantity-based distortions more broadly—are pervasive in markets such as housing, labor, and trade. Policies such as import quotas, visa caps, zoning restrictions, emissions limits, and local content requirements restrict quantities of activities or inputs. Likewise, missing markets constrain quantities of transactions: incomplete credit markets limit transactions across time periods, and incomplete insurance markets limit transactions across states of nature. In this paper, we develop a framework for analyzing quotas and other quantity constraints in economies with general production functions, input-output linkages, and any number of factors and goods.

Much like implicit taxes/wedges, quotas can be used to decentralize distorted allocations. That is, we show that any distorted allocation can be implemented as the competitive equilibrium of an economy with quotas. However, unlike economies with wedges, the resulting equilibrium with quotas is constrained efficient: resources are allocated to maximize output subject to the quotas. Because these economies are constrained efficient, they obey macro-envelope conditions and are not subject to the theory of second best. This greatly simplifies comparative statics.[1]

Using these macro-envelope conditions, we provide three sets of results. First, we provide non-parametric formulas for the effect of changes in quotas and productivity shocks on aggregate output. Second, we characterize nonlinearities in the effects of quota changes on output. Finally, we derive expressions for the misallocation cost of quota distortions—i.e., the loss in output relative to the efficient frontier. We discuss each of these results in turn.

To a first-order, the effect of changing a quota on output is summarized by the profits of producers who own the rights to operate under the quota. Intuitively, because the economy is constrained efficient, the profits earned by quota holders precisely reflect the marginal value of rights to engage in the restricted activity. If a quota does not bind, quota holders earn zero profits, and adjusting the quota has no first-order effect on output. However, when a quota is binding, quota holders earn strictly positive profits, and loosening the quota leads to a first-order improvement in aggregate output.

Likewise, the elasticity of output with respect to a productivity shock is proportional to the affected producer's initial sales less the profits of quota holders. When profits are zero, the effect of productivity shocks is given by the sales of the affected producer, as in

[1]Our baseline framework abstracts from rent-seeking, whereby resources are destroyed as a result of competition for rents, à la Bhagwati (1965) or Krueger (1974). Rent-seeking can occur with either taxes or quotas (see Liu 2019 for an example with taxes). We extend our framework to allow for rent-seeking in Online Appendix B.

Hulten's (1978) theorem. When profits are positive, the effect of productivity shocks is dampened relative to Hulten's theorem because the resources saved from an increase in productivity are diverted to lower marginal value users.

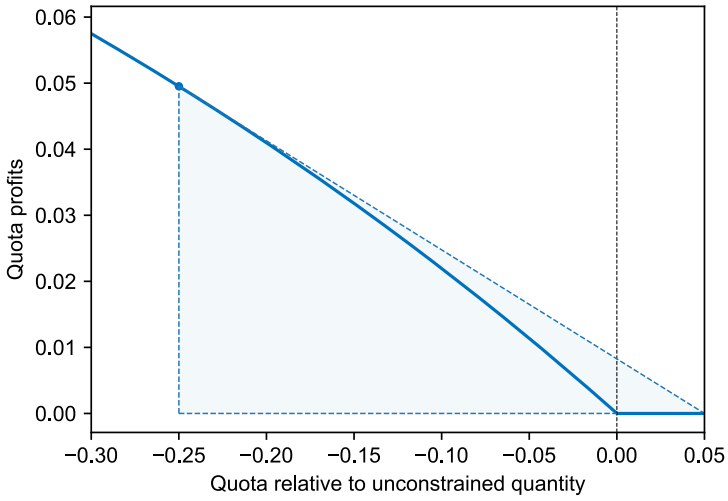
An advantage of working with quotas is that only a few statistics—the profits of quota holders and firms' sales—are needed to characterize these effects. This parsimony arises because of constrained efficiency. When a quota is relaxed, resources flow to the constrained producer from unconstrained uses. Thus, relaxing one quota distortion, holding the rest fixed, always increases output, and the gap between the marginal social value of resources for constrained and unconstrained users is measured by the profit margin of quota holders. Contrast this with economies that instead feature tax- or wedge-like distortions. In such economies, cutting one tax potentially reallocates resources throughout the entire economy, including from other producers that may be underproducing even more than the producer whose taxes are being cut. Thus, whereas analyzing wedge distortions potentially requires rich information about the entire input-output structure of the economy, the underlying elasticities of substitution in production and consumption, and wedges throughout, these issues can be avoided entirely when primitive distortions take the form of quotas.

While these results characterize the effects of marginal changes in quotas, evaluating the effects of larger policy reforms, such as major liberalizations, requires understanding nonlinearities. Since the first-order effect of a quota change depends on the profits of quota holders, the strength of nonlinearities depends on how profits change as the quota is relaxed or tightened. The response of profits to quota changes is therefore a sufficient statistic for nonlinear effects. The elasticity of profits to quantities can either be estimated directly, if exogenous variation exists, or can be constructed from the input-output table and microeconomic elasticities of substitution. One especially tractable case is when a quota is set to maximize the real profits it generates, conditional on other quotas. In this case, the second-order effect of a change in a quota is described by the first-order effect squared.

Interestingly, whereas output is always log concave with respect to quota changes near the efficient point, output can become log convex when the equilibrium is far from the efficient allocation. That is, in economies with preexisting distortions, nonlinearities can amplify the benefits of large liberalizations relative to small ones, and mitigate the losses from further distortions. Such nonlinearities can be important for evaluating proposed policy reforms.

Finally, we characterize the costs of misallocation caused by quota distortions. The key insight is the following. If we know how profits respond to changes in quotas on the

Figure 1: Estimating the distance to the frontier due to a quota distortion.



margin, then we can estimate how much the quota must be relaxed to reach the efficient frontier. Figure 1 illustrates this graphically in a stylized example. This figure shows how profits respond to changes in quotas. Extrapolating linearly, we can estimate the change in the quota needed to reach efficiency (conditional on all other quotas) where profits are zero. This point is precisely the level at which the quota ceases to bind. Thus, to a second order, the gains from removing the quota are approximated by the area of the shaded triangle in Figure 1.[2]

In an economy with multiple quotas, the costs of misallocation are determined by initial quota profits and a *quota demand system* that describes how each quota's price (or profits) responds to changes in every other quota. This system captures how much any individual quota would need to be relaxed to cease binding, as well as interactions between quotas, which depend on how the profits earned by holders of one quota change when another quota is relaxed or tightened. Once this matrix of elasticities is estimated, it can be used to calculate the gains from removing a single quota, any subset of quotas, or eliminating quotas altogether to achieve the first-best allocation.

We demonstrate the applicability of our framework in several empirical examples. Specifically, we explore:

1. How would increasing the cap on H-1B visas affect aggregate output and U.S. GDP?

[2]Falvey (1979), Anderson (1985), and Boorstein and Feenstra (1991) emphasize that industry-level quotas can distort the consumption choices of households across varieties within an industry by causing relative prices to change. For example, higher quality varieties, which have higher prices, experience a smaller proportional increase in their price relative to lower quality, lower price varieties, when industry output is subjected to a quota. This type of misallocation arises endogenously in our framework and is captured by our formulas for the aggregate cost of quotas.

2. What are the gains from loosening zoning restrictions on single-family housing?

3. How costly are Argentina's restrictions on capital outflows?

4. How would the gains from phasing out a subset of U.S. quotas on Chinese textile and clothing exports have compared to the gains from removing all quotas?

5. How costly is the restriction on taxicab medallions in New York City, and to what extent has the entry of ride-share companies in New York relaxed this constraint?

Each of these examples pertains to policies that directly regulate quantities. Our framework allows us to provide (approximate) answers to each question while imposing little structure on the rest of the economy. Our estimates are guided by moments from the data, such as quota prices from auctions or rental markets, or sufficient statistics from natural experiments and micro-data.

The outline of the paper is as follows. Section 2 sets up the framework and presents our results on implementing any feasible allocation with quotas. Section 3 characterizes the first-order effects of quota and productivity changes on output, and Section 4 characterizes nonlinear effects of quota changes. Section 5 presents results on the distance to the efficient frontier. We illustrate how to apply our results in several empirical examples in Section 6. Section 7 describes extensions of our framework developed in the Online Appendices, and Section 8 concludes.

Related literature. This paper is related to a large literature on the causes and costs of misallocation. The classic approach, dating back to Harberger (1954), models misallocation using wedges. The wedge approach has been successfully applied across a range of domains, such as growth accounting (Basu and Fernald 2002), analyzing the drivers of business cycles (Chari et al. 2007), explaining cross-country income differences (Restuccia and Rogerson 2008; Hsieh and Klenow 2009), productivity measurement (Petrin and Levinsohn 2012), calculating social losses from financial frictions and market power (Bigio and La'O 2020; Peters 2020; Edmond et al. 2023), estimating the benefits of reform and liberalization (De Loecker et al. 2016; Bau and Matray 2023), and analyzing monetary non-neutrality (Rubbo 2023). Baqaee and Farhi (2020) provide a general characterization of the efficiency losses from wedge distortions. Our paper provides an analogous characterization of efficiency losses when distortions take the form of quotas rather than wedges.

This paper is also related to a literature that studies how microeconomic shocks affect aggregate efficiency, dating back to Domar (1961) and Hulten (1978). Carvalho and

Tahbaz-Salehi (2019) and Baqaee and Rubbo (2023) survey recent studies taking this approach. This literature can be divided into two branches. One branch focuses on how micro shocks affect aggregate output in efficient economies, for example, Foerster et al. (2011), Gabaix (2011), Acemoglu et al. (2012), Atalay (2017) and Baqaee and Farhi (2019). The other branch emphasizes the importance of inefficiencies, like Baqaee (2018), Grassi (2017), Liu (2019), Reischer (2019), and Buera and Trachter (2024). Our paper is at the intersection of these two branches, since the economies we study feature distortions but are constrained efficient.

Since we focus on quantity-based distortions, our paper is also related to studies that examine the costs of specific quantity-based constraints using quantitative models. For example, Feenstra (1988) estimates the cost of import quotas on Japanese automobiles, and Feenstra (1992) surveys evidence on losses from import quotas and other protectionary trade measures across a wider array of categories. Khandelwal et al. (2013) estimate the costs of quotas on Chinese textile and clothing imports. Other studies estimate the costs of misallocation induced by constraints on housing supply (see e.g., Glaeser and Gyourko 2018; Hsieh and Moretti 2019). We illustrate our sufficient statistics methodology using some of these examples.[3]

2 Framework

In this section, we set up our framework, define an equilibrium with quotas, and show that any feasible allocation can be implemented using quotas. We then demonstrate how an economy with wedges can be mapped to an economy with quotas.

2.1 Setup

Output is the maximizer of a constant-returns aggregator of final demand for goods $1, ..., N$,

$$Y = \max_{\{c_1,...,c_N\}} \mathcal{D}(c_1, ..., c_N),$$

subject to the budget constraint,

$$\sum_{i}^{N} p_i c_i = \sum_{f=1}^{F} w_f L_f + \sum_{i=1}^{N} \Pi_i,$$

[3]Our paper is not closely related to the public finance literature that studies whether policymakers should use quotas or taxes to achieve policy objectives, like raising revenues, under uncertainty (see, for example, Weitzman 1974 or Dasgupta and Stiglitz 1977).

where c_i is the representative household's final demand for good i, p_i is its price, w_f is the wage of factor f, L_f is the supply of factor f, and Π_i are the profits earned by producers of good i. We require that all final demands c_i are non-negative and assume that $\mathcal{D}$ is weakly increasing in each argument. We take nominal output as the numeraire, $\sum_i^N p_i c_i = 1$.

Each good i is produced by competitive firms using the production function,

$$y_i = A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ..., L_{iF}),$$

where x_{ij} is the quantity of good j used in the production of good i, L_{if} is the quantity of factor f used by i, and A_i is a Hicks-neutral productivity shifter. We assume that F_i is constant returns to scale and weakly increasing in each argument, and we require that all inputs x_{ij} and L_{if} are non-negative. We denote the Domar weight of good i by $\lambda_i = p_i y_i$, and the Domar weight of factor f by $\Lambda_f = w_f L_f$.

A *quota* restricts the output of good i at a quantity y_{i^*},

$$y_i = \min\{y_{i^*}, A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ..., L_{iF})\}.$$

Profits for producers of good i are total revenues less costs of intermediate inputs and factors,

$$\Pi_i = p_i y_i - \sum_{i=1}^{N} p_j x_{ij} - \sum_{f=1}^{F} w_f L_{if}.$$

As anticipated by the representative household's budget constraint, profits of all producers are rebated to households lump sum. Since each production function F_i is constant returns to scale, equilibrium profits in the absence of quotas are zero, but may be strictly positive when quotas are binding.

Resource constraints for each good $1 \leq i \leq N$ and each factor $1 \leq f \leq F$ are

$$c_i + \sum_{j=1}^{N} x_{ji} \leq y_i \qquad \text{and} \qquad \sum_{i=1}^{N} L_{if} \leq L_f.$$

Definition 1 (Equilibrium with quotas)**.** Given quotas y_{i^*}, productivities A_i, production functions F_i, and factor supplies L_f, an *equilibrium with quotas* is a set of prices p_i, factor wages w_f, outputs y_i, final demands c_i, and intermediate and factor input choices x_{ij} and L_{if} such that: final demand maximizes the final demand aggregator subject to the budget constraint; each producer maximize profits taking prices as given; $y_i \leq y_{i^*}$ for each good with a quota; and resource constraints for all goods and factors are satisfied.

2.2 Implementing an Allocation Using Quotas

Our first result is that any feasible allocation—i.e., any allocation of resources and intermediate inputs that obeys production technologies and market clearing conditions—can be implemented as a decentralized equilibrium of an economy with quotas.

Definition 2. An allocation $\{c_1, ..., c_N, x_{i1}, ...x_{iN}, L_{i1}, ..., L_{iF}\}_{1 \leq i \leq N}$ is *feasible* if c_i, x_{ij}, and L_{if} are all non-negative, $y_i = A_i F_i(x_{i1}, ..., x_{iN}, L_{i1}, ...L_{iF})$ for all i, and the resource constraints hold: $c_i + \sum_{j=0}^{N} x_{ji} \leq y_i$ for all i and $\sum_{i=1}^{N} L_{if} \leq L_f$ for all f.

Proposition 1 (Implementation via implicit quotas)**.** *Suppose an allocation $\mathcal{X}$ is feasible. Then there exists an economy with quotas in which the allocation of the decentralized equilibrium is $\mathcal{X}$. Moreover, given these quotas, the allocation $\mathcal{X}$ is efficient.*

The intuition for Proposition 1 comes from the fact that by introducing additional producers and using quotas, one can guarantee that the competitive equilibrium yields any desired feasible allocation. First, to ensure that the use of good j in the production of i is equal to x_{ij}, one can create a new producer k such that i's use of j flows through k. Then, introducing a quota on the output of good k at $y_{k^*} = x_{ij}$ guarantees that the use of good j by i is at most x_{ij}. Further quotas on every other use of good j, combined with the fact that the final demand aggregator is increasing in all goods, can also guarantee that the use of good j by i is at least x_{ij}. Thus, given these quotas, the decentralized equilibrium with competitive firms yields exactly the desired allocation.

Since the allocation can be decentralized as the equilibrium of the competitive economy, the first welfare theorem also implies that the allocation is constrained efficient. That is, the allocation $\mathcal{X}$ maximizes output among the set of allocations in the production possibilities frontier of the economy with quotas.

We show how quotas can implement any feasible allocation in a stylized model of a small open economy. We will return to this example to illustrate several results throughout the paper.

Example 1 (Small Open Economy)**.** Consider a small open economy in which labor is the sole domestic factor and is used to produce a domestic good, denoted by d. Import-export firms, denoted by m, trade the domestic good for a foreign good (f) and sell the foreign good to households. The exchange rate between the domestic good and the foreign good is fixed at an exogenous price p_m, and there is an iceberg trade cost τ, so that τp_m units of the domestic good must be exchanged to import one unit of the foreign good. We

impose that trade is balanced.[4] Household welfare is given by the constant elasticity of substitution (CES) aggregate,

$$Y = \left(\omega c_d^{\frac{\theta-1}{\theta}} + (1-\omega) c_f^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}},$$

where c_d and c_f are household consumption of the domestic and foreign goods, θ is the trade elasticity, and ω is a taste shifter that determines the degree of home bias.

The set of feasible allocations in this economy is $\{(y_d, c_d, c_f) \in \mathbb{R}^3_+ \mid \tau p_m c_f + c_d \leq y_d \leq L\}$. We can implement any allocation in this set by introducing three quotas: a quota on labor supply, a quota that caps imports of the foreign good by import-export firms, and a quota on the consumption of the domestic good. These quotas ensure that the total amount of domestic good produced is y_d, that no quantity of the domestic good in excess of $c_f/\tau p_m$ is used for trade, and that no quantity in excess of c_d is used for consumption. Given these three constraints, the decentralized equilibrium of the economy with quotas has the allocation (y_d, c_d, c_f). Moreover, since real output is strictly increasing in c_d and c_f, it follows immediately that an allocation in which c_d and c_f are equal to the respective quotas on domestic-good consumption and imports maximizes output subject to the quota constraints. Thus, quotas implement any desired feasible allocation in this economy, and the resulting allocation always maximizes output subject to the quota constraints.

2.3 Mapping Wedges to Quotas

The classic approach to modeling misallocation uses implicit taxes, or "wedges," to implement distorted allocations of resources.

Definition 3 (Equilibrium with wedges)**.** Given wedges τ_i, productivities A_i, production functions F_i, and factor supplies L_f, an *equilibrium with wedges* is a set of prices p_i, factor wages w_f, outputs y_i, final demands c_i, and intermediate and factor input choices x_{ij} and L_{if} such that: final demand maximizes the final demand aggregator subject to the budget constraint; each producer minimizes costs taking prices as given; the price of good i equals its marginal cost times the exogenous wedge τ_i; wedge revenues $\Pi_i = (1 - 1/\tau_i)\, p_i y_i$ are rebated to the representative household; and resource constraints for all goods and factors are satisfied.[5]

[4] While our general model is of a closed economy, rather than an open economy, this stylized model of a small open economy is a special case of our framework since balanced trade with exogenous terms-of-trade is equivalent to having a linear technology that converts domestic goods into foreign goods with some exogenous rate of transformation.

[5] The assumption that wedges are applied to output prices is without loss of generality, since user-good-specific can be modeled by introducing intermediaries with wedges.

A challenge when comparing economies with quotas and wedges is that two economies that share the same physical allocation of resources, when implemented via implicit quotas or wedges, may have different prices, sales shares, and profits. This challenge stems from the fact that it is often possible to implement a given allocation of resources with many different sets of wedges. To take an example, consider a horizontal economy in which firms use labor to produce differentiated varieties, which are then consumed by a representative household. In this economy, doubling all firms' markups increases firms' prices and profits and reduces labor's share of income without affecting the allocation of resources.

We can eliminate this indeterminacy by imposing restrictions on wedges. Proposition 2 presents restrictions that ensure that if the allocation of resources in a wedge economy coincides with a quota economy, the observable prices, sales, and profits also coincide.

Proposition 2 (Matching observables in wedge and quota economies)**.** *Consider an economy with quotas in which all producer prices p_i and factor wages w_f are strictly positive. Consider a second economy in which the same allocation of resources is implemented with wedges τ_i on each firm. If $\tau_i \geq 1$ for all i and, for each good or factor i, either the good is directly consumed by the household $c_i > 0$ or there exists some producer j such that $\partial F_j / \partial x_{ji} > 0$ and $\tau_j = 1$, then prices, sales, and profits are identical across the two economies.*

The first condition that $\tau_i \geq 1$ for all producers ensures that profits in the wedge economy are weakly positive. This is necessary to match observables across the wedge and quota economies, because quota profits must be weakly positive (they are strictly positive when quotas are binding or else zero).

The second condition requires that one user of each factor or good in the economy (which may be the representative household) has a wedge $\tau_i = 1$. In an economy with quotas, if all users of a good have binding quotas, the price of that good must be equal to zero. The assumption that all prices and wages in the economy with quotas are strictly positive thus implies that at least one user of each factor or good must be unconstrained.

Together, the first and second conditions also ensure that the wedges that map to a given quota allocation are unique. Since all producers must have weakly positive profits, and at least one producers' profits must be exactly zero among users of each good, one cannot scale up wedges across firms while continuing to satisfy these requirements. Thus, the conditions in Proposition 2 identify the unique vector of wedges that generate the same allocation and prices as a given set of quotas.

Example 2 (Small Open Economy)**.** Consider the small open economy from Example 1, and suppose the only binding quota is the quota on imports y_{m^*}. Suppose we have an

identical economy (the "tariff economy") where, instead of an import quota, there is an import tariff t_m and a tax on consumption of the domestic good t_d. Given total production of the domestic good y_d and the domestic-good consumption tax t_d, the tariff t_m that implements the same import quantity y_{m^*} as the economy with quotas is

$$t_m = \frac{1-\omega}{\omega} \frac{t_d}{\tau p_m} \left(\frac{y_d - y_{m^*} \tau p_m}{y_{m^*}} \right)^{\frac{1}{\theta}}. \tag{1}$$

Notice that the import tariff t_m and the tax on domestic good consumption t_d can be scaled by an arbitrary factor without altering the import quantity.

Setting the tax on the domestic good $t_d = 1$ leads prices, sales, and profits to coincide across the tariff economy and the quota economy. For example, in the quota economy, the government earns profits Π_m by selling export licenses to import-export firms. It is straightforward to verify that in the tariff economy with $t_d = 1$, the same level of profits Π_m accrues to the government as tariff revenue.

3 First-Order Effects

In this section, we characterize the response of output to changes in quotas and productivities up to a first order. Since economies with quotas are constrained efficient, these effects can be expressed non-parametrically in terms of a small set of sufficient statistics.

3.1 First-Order Effects of Quota Changes and Productivity Shocks

Proposition 3 describes the change in output resulting from changes to quotas and producer productivities.

Proposition 3 (First-order effects)**.** *To a first order, the change in output resulting from changes in quotas y_{i^*} and productivities A_i is*

$$d \log Y = \sum_{i^*} \Pi_i d \log y_{i^*} + \sum_i (\lambda_i - \Pi_i) \, d \log A_i.$$

If all quotas are initially non-binding, then $d \log Y = \sum_i \lambda_i d \log A_i$.

The effect of a change in quota y_{i^*} on output is proportional to the profits of the constrained producers. Positive profits indicate that the quota is a binding constraint on production. Thus, when profits are positive, relaxing the quota constraint on production increases the production of the good and total output. Note that calculating the effect of

relaxing a quota does not require specifying where in the economy the additional resources used in the production of i will come from. Because the economy is constrained efficient, producer i's profits already reflect the value of assigning it more resources relative to unconstrained producers.

Likewise, the effect of a change in productivity on output is given by the sales share of the affected producer minus its profits. If a producer's profits are non-zero, a quota constrains its output. Rather than increasing its output, an increase in the producer's productivity frees up some of the resources that were required to produce the amount of output given by the quota. The contribution of those freed-up resources to output is exactly equal to their sales the economy, i.e., the costs of the constrained producer on those resources.

In an economy without any quota distortions, all profits are zero. In this case, Proposition 3 shows that the comparative statics converge to familiar results from efficient economies. Specifically, the introduction of marginal distortions has no first-order effect on output, since efficiency equates the marginal benefit of inputs across all uses. The elasticity of output to productivity shocks is exactly equal to the sales shares of affected producers, as in Hulten's (1978) theorem.

We illustrate these results in the small open economy from Example 1.

Example 3 (Small Open Economy)**.** Consider the small open economy from Example 1, and suppose the only binding quota is the import quota y_{m^*}. We apply Proposition 3 to see how a change to the import quota and a change in iceberg trade costs affect output.

The effect of a change in the import quota by $d \log y_{m^*}$ is:

$$\frac{d \log Y}{d \log y_{m^*}} = \Pi_m. \tag{2}$$

That is, the output gains from increasing the import quota are proportional to the profits of import-export firms. If the quota does not bind, perfectly competitive import-export firms make zero profit, and further increases in the import quota have no effect on output. On the other hand, when the import quota binds and import-export firms make positive profits, relaxing the quota results in output gains.

Changes to the iceberg trade costs τ are equivalent to increasing the productivity of import-export firms in exchanging the domestic good for the foreign good. We can therefore use Proposition 3 to calculate the output gains from reducing trade costs by $-d \log \tau$:

$$-\frac{d \log Y}{d \log \tau} = \lambda_f - \Pi_m = (1 - \lambda_f)\left(\frac{y_d - c_d}{c_d}\right). \tag{3}$$

Due to the import quota, a reduction in trade costs does not actually increase household consumption of imported goods. But it does reduce the amount of domestic good that is required for exchange with the foreign good. As a result, the reduction in trade costs increases output by increasing the quantity of domestic good that remains for consumption by households. Thus, the output gains from a reduction in trade costs are also equal to the household expenditure share on the domestic good, multiplied by the ratio of the amount of the domestic good used for trade to the amount consumed.

3.2 Comparison to Economies with Wedge Distortions

It is useful to contrast the effect of shocks in an economy with quotas to the effect of shocks in an economy with wedge distortions.

Proposition 4 (Comparative statics in economies with wedge distortions)**.** *In an economy with wedge distortions, the effect of wedge shocks* $d \log \tau_i$ *and productivity shocks* $d \log A_i$ *on output is*

$$d \log Y = \sum_i \sum_j \Pi_i \left[\frac{\partial \log y_i}{\partial \log \tau_j} d \log \tau_j + \frac{\partial \log y_i}{\partial \log A_j} d \log A_j \right] + \sum_i (\lambda_i - \Pi_i) \, d \log A_i, \qquad (4)$$

where $\partial \log y_i / \partial \log \tau_j$ *and* $\partial \log y_i / \partial \log A_j$ *are general-equilibrium elasticities of* y_i *with respect to changes in* τ_j *and* A_j *respectively.*

As in the case of an economy with quotas, if profits for all producers are initially zero, marginal wedge distortions have no effect on output, and the effect of productivity shocks are given by Hulten's theorem. However, if there are existing distortions, the effect of wedge shocks and productivity shocks on output depends on how wedge and productivity changes affect all producers' quantities. Computing these effects generally requires information on the input-output matrix and elasticities of substitution in production and consumption. Moreover, in economies with multiple wedge distortions, there is no guarantee that removing the wedge on one producer will improve efficiency and output, due to the theory of second best (Lipsey and Lancaster 1956).

The usefulness of Proposition 3 over Proposition 4 depends on the extent to which quotas can be treated as primitives. If the mapping from primitive shocks to changes in quotas is itself complicated, then Proposition 3 is less useful. For example, if the primitive economy features taxes, and we represent that allocation using quotas instead, then all quotas may need to move in response to changes in taxes. In this case, calculating the endogenous changes in quotas ultimately requires the same information about the

structure of the economy as is required to calculate the effects of changes in wedges (e.g. information about the input-output structure, elasticities of substitution, returns to scale, and so on). However, in cases where the primitive distortions are quota-like, then the equivalent wedge-representation in (4) is complex and requires assumptions about the structure of the economy that, given Proposition 3, are unnecessary.

Example 4 (Small Open Economy)**.** We compare the effect of shocks in the small open economy when the allocation is implemented with an import tariff rather than an import quota. First, consider the effect of a change in the import tariff $d \log t_m$ on output. Log-linearizing the expression for the tariff in (1) and applying Proposition 4, we find

$$\frac{d \log Y}{d \log t_m} = \Pi_m \frac{d \log y_{m^*}}{d \log t_m} = -\theta \Pi_m \frac{c_d}{y_d}.$$

Increases in the tariff reduce output. The effect is stronger when the trade elasticity, θ, is high because a higher trade elasticity results in a greater reduction in imports. The effect is also stronger when profits generated by the tariff, Π, are high because this indicates a larger initial distortion. Finally, the effect is also stronger when the economy is more open, as measured by the ratio of the domestic good used for domestic consumption.

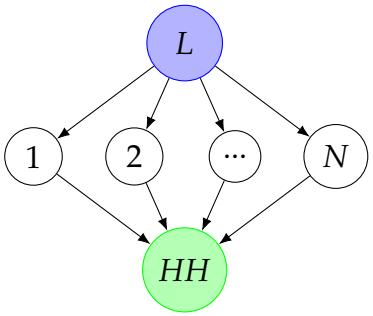
Note that, unlike the effect of changes to the import quota in (2), calculating the effect of tariff changes on output requires knowing both the trade elasticity θ and information about the economy's structure (in this case, the share of domestic good used for consumption, c_d/y_d). This distinction highlights a broader difference between working with quotas and tariffs: calculating the output effects of quota changes requires only observable profits, while calculating the effects of tariff changes requires more detailed knowledge of the underlying economic structure, such as elasticities of substitution in consumption and production, to translate wedge changes into the quantity changes that affect output.

Likewise, the effect of a decline in trade costs in the tariff economy is:

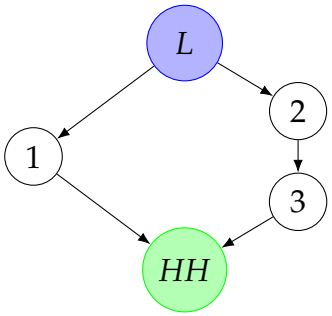
$$-\frac{d \log Y}{d \log \tau} = \lambda_f - \Pi_m + \frac{\Pi_m}{1 - \Pi_m} \left[\left(\lambda_f - \Pi_m \right) + \theta \left(1 - \lambda_f \right) \right].$$

This expression coincides with (3) when the import tariff is zero and the import quota is not binding (i.e., $\Pi_m = 0$), but otherwise differs since a reduction in trade costs in the tariff economy generally increases the quantity of imports. In other words, despite the two economies sharing the same initial allocation of resources, the effect of changes in trade costs across the two economies generally differs depending on whether the primitive distortion takes the form of a quota or tax. Note that computing the effect of the decline in trade costs in the tariff economy again requires knowing the trade elasticity θ, which is

Figure 2: Illustrative examples.



(a) Horizontal economy.



(b) Interdependent producers.

not necessary to compute the effect of the decline in trade costs in the quota economy.

The previous example highlights that whether distortions take the form of quotas or wedges matters for the reallocations that take place in response to shocks. In economies with quotas, when a quota is relaxed, resources are reallocated to a constrained producer from other, unconstrained uses. In contrast, in economies with wedge distortions, reducing the wedge on one producer reallocates resources throughout the economy, even parts of the economy that are more constrained than the producer whose wedge is reduced.

We illustrate how the reallocations triggered by a reduction in a quota versus a wedge differ in the following example.

Example 5 (Reallocations Under Quotas vs. Wedges)**.** Consider the horizontal economy illustrated in Figure 2a. Firms $1, ..., N$ use labor to produce varieties. A representative household has CES preferences over these varieties with an elasticity of substitution θ. We compare how relaxing a distortion on firm 1 affects output when distortions are implemented with quotas versus with wedges.

When distortions are implemented with quotas, Proposition 3 describes the effect of relaxing the constraint on firm 1:

$$d \log Y = \Pi_1 d \log y_{1^*}. \qquad \text{(Quota economy)}$$

The effect of relaxing the quota is always weakly positive and is strictly positive if the quota was initially binding (i.e., $\Pi_1 > 0$). The output of any other firms in the horizontal economy with binding quotas is unchanged to a first-order, and so the resources that are reallocated to firm 1 as the quota is relaxed come only from initially unconstrained firms. These unconstrained firms are precisely those where the marginal benefit of resources is initially lowest, and so the reallocation of resources from them toward firm 1 always weakly improves output.

If the same allocation were instead implemented with wedges, we can apply Proposition 4 to calculate the effect of a change in the wedge on firm 1 that increases 1's output by $d \log y_1$:

$$d \log Y = \Pi_1 d \log y_1 - \frac{l_1}{1 - l_1} \left(\sum_{i \neq 1} \Pi_i \right) d \log y_1, \qquad \text{(Wedge economy)}$$

where $l_1 = L_1/L$ is the share of labor used by firm 1. The effect of relaxing the distortion on firm 1 in the wedge economy differs from the quota economy because as the wedge on firm 1 falls, labor is reallocated to firm 1 proportionately from all other firms. In other words, the resources gained from firm 1 no longer come only from unconstrained firms, but instead from the cross-section of all other firms.

Since resources are reallocated proportionately from other firms, the effect of reducing the wedge on firm 1 depends on how binding the constraint is on firm 1 relative to the average firm. Even when $\Pi_1 > 0$, it is possible for firm 1 to be less constrained than the average firm in the horizontal economy, and for the overall output effect to thus be negative. In other words, the presence of multiple distortions in the second best means that reducing the the extent of one distortion may actually exacerbate other distortions and reduce, rather than improve, efficiency (Lipsey and Lancaster 1956).

Analyzing the effect of shocks in economies with wedge distortions is further complicated by interdependencies across firms. When an allocation is implemented with wedges, a comparison of wedges across producers is generally not sufficient to assess whether relaxing the distortion on a firm improves efficiency. In contrast, economies with quotas are constrained efficient, and removing a quota always improves aggregate efficiency.

Example 6 (Interdependent Producers)**.** Consider the economy in Figure 2b: firm 1 produces a consumption good using labor, firm 2 produces an intermediate that is used by firm 3 to produce a consumption good, and households have CES preferences over the consumption goods produced by firms 1 and 3 with an elasticity of substitution θ.

Suppose first that an allocation of resources in this economy is implemented with wedges τ_1, τ_2, and τ_3. Applying Proposition 4, the effect of reducing the wedge on firm 2 is

$$d \log Y = \sum_i \Pi_i \frac{\partial \log y_i}{\partial \log \tau_2} d \log \tau_2 = \theta \left[\Pi_1 - (\Pi_1 + \Pi_2 + \Pi_3) \frac{L_1}{L_1 + L_2} \right] d \log \tau_2,$$

where L_1 and L_2 are the quantities of labor used by firm 1 and 2. Notice that this effect can be positive or negative depending on firms' initial profits. That is, removing the

distortion does not unambiguously increase output. Moreover, comparing τ_2 to τ_1 alone is not sufficient to identify whether removing the wedge on firm 2 increases output, because of the interdependence between firm 2 and firm 3. The importance of these interdependencies for evaluating policies was emphasized by McKenzie (1951).

If the same allocation is instead implemented with quotas, Proposition 3 shows that the effect of relaxing the quota on firm 2 is instead

$$d \log Y = \Pi_2 d \log y_{2^*}.$$

Removing the quota distortion always increases output, and the profits of firm 2 can be used to estimate the benefits of relaxing the quota without having to take into account interdependencies across producers.

4 Nonlinearities

Evaluating the effect of larger policy reforms, such as major liberalizations, requires understanding potential nonlinearities that may arise due to large shocks. In this section, we characterize the response of output to a change in quotas to a second order. These nonlinearities dictate how the effects of large shocks differ from small shocks.

4.1 Nonlinear Effects of Quotas

Proposition 5 characterizes the response of output to changes in quotas to a second order.

Proposition 5 (Nonlinear effects of quotas)**.** *The effect of a vector of quota changes* $\mathbf{d} \log \mathbf{y}_*$ *on output to a second order is*

$$\Delta \log Y \approx \mathbf{\Pi}' \mathbf{d} \log \mathbf{y}_* + \frac{1}{2} \left(\mathbf{d} \log \mathbf{y}_* \right)' H \left(\mathbf{d} \log \mathbf{y}_* \right),$$

where H *is a symmetric matrix with* $H_{ij} = \partial \Pi_i / \partial \log y_{j^*}$ *equal to the semi-elasticity of profits* Π_i *to changes in the quota on producer* j*. When there is a change to only a single quota* y_{i^*}*, the effect on output to a second order is*

$$\Delta \log Y \approx \Pi_i d \log y_{i^*} + \frac{1}{2} \frac{d \Pi_i}{d \log y_{i^*}} (d \log y_{i^*})^2.$$

Since the effect of a change in a quota to a first order depends on the profits of the constrained firms, the nonlinear effects depend on how profits of the constrained firms

change as the quota changes. For example, if tightening a quota leads to a rise in profits, then rising profits amplify the output losses that result from a large reduction in the quota level. Conversely, if profits fall when the quota is tightened, then nonlinear effects partially mitigate the output losses from a large shock.

When there are changes to multiple quotas, how profits of all quotas change in response to variation in the quota levels is summarized by the matrix H. The matrix H is a *quota demand system*: each entry captures how quota prices (and profits) respond to changes in quota levels. A useful feature of this matrix H is that it is symmetric. This property stems from the fact that profits are the elasticity of output with respect to quota changes, $\Pi_i = \partial \log Y / \partial \log y_{i^*}$ (see Proposition 3).[6] Thus, the matrix of semi-elasticities of profits with respect to quota changes is the Hessian of output with respect to quotas. In practice, the symmetry of H also reduces the number of empirical moments needed to estimate the full matrix.

While the matrix H can be estimated using exogenous variation in the data, Appendix Proposition A1 shows how one can compute each of the semi-elasticities $H_{ij} = d\Pi_i / d \log y_{j^*}$ in terms of the structure of the input-output network, producer sales shares, and microeconomic elasticities of substitution. These results exploit the fact that a quota can be recast as a factor with fixed supply y_{i^*}, where profits Π_i are equivalent to the factor's share of income and changes to the quota are equivalent to shocks to the factor's productivity. Thus, the response of factor income shares to productivity shocks in efficient economies (characterized in Baqaee and Farhi 2019) also describes the response of profits to quota changes.

Around the first-best allocation where resource use is unconstrained by quotas, output is always log concave with respect to quota changes. Away from the efficient point, however, output may be log convex with respect to changes in quotas. When output is log convex with respect to quota changes, nonlinearities mitigate the downsides of further restricting output and amplify the benefits of loosening quotas. In other words, large restrictions that curtail an activity are relatively less costly than small quantity reductions, and large liberalizations that relax quotas are more beneficial than incremental ones.[7] We illustrate this in an example of horizontal economy.

Example 7 (Horizontal Economy with a Single Quota)**.** Consider the horizontal economy in Figure 2a. We consider how nonlinearities shape the effect of a change in the quota on

[6]The symmetry of H is a consequence of the fact that final demand maximizes a homothetic aggregator. If final demand is not homothetic, H needs to be adjusted to account for income effects (see Baqaee and Burstein 2023).

[7]These statements characterize nonlinearities in terms of log changes in quotas, $d \log y_{i^*}$. The concavity or convexity of output in terms of changes in quota *levels*, dy_{i^*}, may differ due to Jensen's inequality.

a firm i on output.

Applying Proposition 5 to this economy, and using the explicit characterization of H in Appendix Proposition A1, we find that the response of output to a change in the quota on firm i is

$$\Delta \log Y \approx \Pi_i d \log y_{i^*} + \frac{1}{2}\left[\frac{\Pi_i}{1-\Pi_i} - \frac{1}{\theta}\frac{\lambda_i}{1-\lambda_i}\right](1-\Pi_i)^2 \left(d \log y_{i^*}\right)^2. \tag{5}$$

The first term in (5) reflects the first-order effect of quota changes on output and is familiar from Proposition 3. The second term in (5) reflects nonlinearities in how changes in the quota on firm i affects output and, as seen in Proposition 5, depends on how the firm's profits evolve as the quota changes.

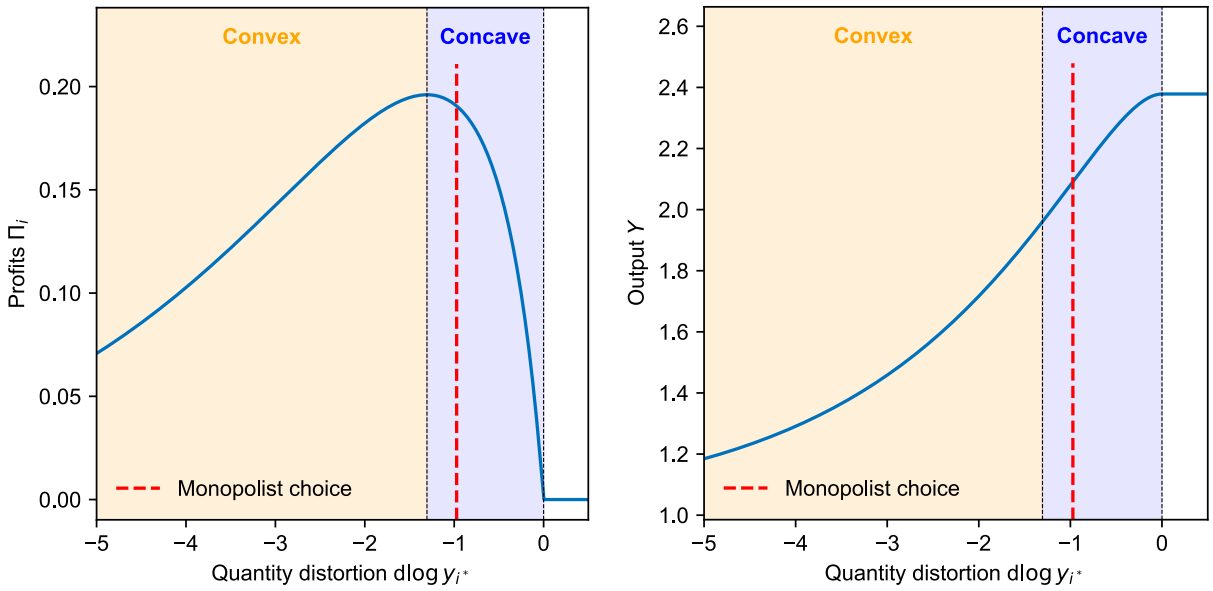
The sign of this second-order term depends on the initial level of profits, Π_i, as well as the household's elasticity of substitution θ and firm's initial sales share λ_i. Close to efficiency, this term is negative because $\Pi_i \approx 0$, meaning that output is concave in quota changes: nonlinearities exacerbate the effects of negative shocks and dampen the effects of positive shocks. However, away from the efficient point, the second-order term may be positive if $\theta > 1$. A positive second-order term implies that an increase in the quota increases profits. When this is the case, nonlinearities amplify the benefits of positive shocks and mitigate further losses from negative shocks.

Figure 3 illustrates these results in a simulation of the horizontal economy with $\theta > 1$. The left panel shows that profits Π_i are hump-shaped in the quota y_{i^*}. Starting at the point where the quota is just binding (i.e., $d \log y_{i^*} = 0$), a decrease in the quota initially increases profits. But when the quota is sufficiently low, further tightening the quota in fact leads profits to decline. The non-monotonic path of profits means that output, shown in the right panel of Figure 3, switches from concave in the region near the efficient point to convex in the quota at points sufficiently far from the efficient frontier.

The changing sign of these nonlinearities means that a comparison of the effects of large and small shocks will differ depending on the initial level of the quota. Suppose the initial level of the quota is in the concave region in Figure 3. Then, the gains from relaxing the quota on firm i will peter out as the change in the quota becomes larger. In other words, the gains from a marginal increase in the quota overstate the gains that would result from a large increase in the quota. Conversely, if the initial level of the quota is sufficiently low, then the gains from a small change to the quota understate the gains that would result from a large liberalization.

In an economy with multiple quotas, interactions between changes in multiple quotas show up as nonlinearities on the effect on output. We show how the matrix H determines

Figure 3: Nonlinearities away from the frontier in a horizontal economy.



(a) Profits.

(b) Output.

Note: The thick dashed line is the output quantity chosen by a monopolist to maximize real profits. Simulation of two identical firms in a horizontal economy with an elasticity of substitution $\theta = 1.8$.

these interactions in the following example.

Example 8 (Horizontal Economy with Multiple Quotas)**.** Consider again the horizontal economy from Figure 2a, and suppose there are changes to the quotas on two firms, y_{1^*} and y_{2^*}. Following Proposition 5, the effect on aggregate output is given by

$$\Delta \log Y \approx \underbrace{\Pi_1 d\log y_{1^*} + \Pi_2 d\log y_{2^*}}_{\text{First order}} + \underbrace{(1/2)\left(H_{11}\left(d\log y_{1^*}\right)^2 + H_{22}\left(d\log y_{2^*}\right)^2 + 2H_{12}\left(d\log y_{1^*}\right)\left(d\log y_{2^*}\right)\right)}_{\text{Second order}}.$$

How the interaction between the two quota changes affects output depends on the sign of H_{12}. When H_{12} is positive, relaxing the quota on one firm increases the profits that accrue to the second quota. Thus, relaxing both quotas together amplifies efficiency gains relative to loosening each quota independently. Conversely, when H_{12} is negative, relaxing one quota makes the second quota less binding, and hence reduces the incremental gains that would be achieved from also relaxing the second quota.

We can solve for the conditions under which $H_{12} \lesseqgtr 0$ using the expressions from

Proposition A1. We find that H_{12} is positive if

$$\theta < 1 - \frac{(\lambda_1 - \Pi_1)(\lambda_2 - \Pi_2)}{\lambda_3 \Pi_1 \Pi_2}.$$

Two insights emerge. First, when the economy is efficient and $\Pi_1 = \Pi_2 = 0$, H_{12} is always negative, and thus the gains from relaxing both quotas around the efficient point is always lower than the sum of the gains from relaxing each quota individually. The intuition is that, when both quotas are just binding, tightening the quota on firm 1 pushes more resources to firm 2 and thus makes the existing quota on firm 2 more restrictive. But the effects of positive profits at both firms can be undone by relaxing the quota solely on firm 1—thus the incremental gains from relaxing both quotas is less than the gains from relaxing each quota individually.

Second, when $\Pi_1, \Pi_2 > 0$, a necessary condition for H_{12} to be positive in this economy is that the firms' outputs are complements ($\theta < 1$). Intuitively, when outputs are complements, an increase in the supply of output by firm 1 increases the marginal value of outputs from firm 2. This force amplifies the gains from relaxing the quotas on both firms together compared to relaxing each individually. When θ is sufficiently low and firms have sufficiently high initial profits, this force can lead the net effect of relaxing both quotas together to be greater than each alone.

We end this section with a special case where it is possible to compute nonlinear effects even without direct knowledge of the input-output matrix and elasticities of substitution. Proposition 6 shows that if a quota is chosen to maximize the real profits it generates, say by a monopolist, then we can characterize nonlinearities in terms of profits alone.

Proposition 6 (Nonlinear effects with a monopolist)**.** *Suppose all production of i is controlled by a monopolist that chooses its output quantity y_i to maximize real profits, taking all other producers' production technologies and quotas as given. Then, the effect of changes in the monopolist's quantity on output to a second order are*

$$\Delta \log Y \approx \Pi_i d \log y_i - \frac{1}{2} \Pi_i^2 (d \log y_i)^2.$$

Output is log concave with respect to changes in the monopolist's output quantity, so nonlinearities always amplify the losses from a further reduction in output by the monopolist and moderate the gains from increasing the monopolist's production. The larger these profits, the faster the gains from increasing the quota peter out relative to the first-order approximation. For example, Figure 3 shows that the quota that maximizes

real profits (indicated by the dashed red lines) in Example 7 is in the region where output is log concave.

5 Distance to the Frontier

In this section, we characterize the misallocation costs of quotas—that is, the output loss relative to the efficient frontier where quota distortions are removed. We provide three non-parametric expressions for the distance to the frontier. These expressions can be used to analyze the effect of relaxing a single quota or relaxing multiple quotas at once.

5.1 Theoretical Results for the Distance to the Frontier

Output Y is maximized when there are no quotas distorting output. Given an initial equilibrium with quotas, Proposition 7 calculates a second-order approximation for the difference in output between the initial equilibrium and the point at which a subset or all of the quota distortions are removed.

Proposition 7 (Distance to the frontier)**.** *Let* $\mathbf{y}_*$ *be a vector of quotas and* $\mathbf{y}^{\mathbf{eff}}$ *be the vector of output quantities that would result if quotas on producers* $i \in \mathcal{I}^*$ *were relaxed to the point of being non-binding. Let* $\mathbf{\Pi}(\mathbf{y}_*)$ *be vector of producers' profits given quotas* $\mathbf{y}_*$*, and define the vector of quantity distortions* $\mathbf{d}\log\mathbf{y}_* = \log\mathbf{y}_* - \log\mathbf{y}^{\mathbf{eff}}$*. For small quantity distortions, the output gains from relaxing all quotas* $i \in \mathcal{I}$ *up to a second order in the quantity distortions* $\mathbf{d}\log\mathbf{y}_*$ *is*

$$\Delta \log Y \approx -\frac{1}{2}\mathbf{\Pi}'\mathbf{d}\log\mathbf{y}_*. \tag{6}$$

Equivalently,

$$\Delta \log Y \approx -\frac{1}{2}\left(\mathbf{d}\log\mathbf{y}_*\right)' H \left(\mathbf{d}\log\mathbf{y}_*\right), \tag{7}$$

where H *is a symmetric matrix with* $H_{ij} = \partial\Pi_i/\partial \log y_{j^*}$ *equal to the semi-elasticity of profits* Π_i *to changes in the quota on producer* j*. Finally, the output gains to a second order in quantity distortions* $d\log y_i$ *can also be calculated using*

$$\Delta \log Y \approx -\frac{1}{2}\mathbf{\Pi}' H^{-1}\mathbf{\Pi}. \tag{8}$$

For (7) *and* (8)*, the matrix* H *can be evaluated either at the equilibrium with quotas or at the equilibrium where all quotas in* $\mathcal{I}^*$ *are relaxed.*

Equation (6) expresses the distance to the frontier in terms of profits and the size of

quantity distortions. When distortions are small, the effect of removing distortions to a second order can be calculated by averaging the first-order effect of changing quotas at the initial equilibrium, given by Proposition 3, and the first-order effect of changing quotas at the efficient point, which is zero by the Envelope theorem. Notice the distance to the frontier in (6) can be computed without requiring additional information about elasticities of substitution in production or other details about the structure of the economy, which are typically required to measure the distance to the frontier when there are wedge distortions (see e.g., Baqaee and Farhi 2020).

Alternatively, profits close to the efficient point can also be estimated by specializing the nonlinear effects from Proposition 5 to an economy that is initially efficient. Since profits at the efficient point are zero, the first-order term disappears, and we are left with (7). The matrix H, which captures the response of profits on each quota to changes in other quotas, describes the misallocation cost of a vector of quantity distortions. We note that, for the second-order approximation in (7), these semi-elasticities can be calculated at either the efficient point or at the observed inefficient allocation.

Both expressions in (6) and (7) requiring knowing the size of quantity distortions $\mathbf{d}\log\mathbf{y}_*$, or equivalently, the output quantities that would prevail if there were no quotas. For cases where it is difficult to ascertain the size of quantity distortions, Equation (8) provides a formula for the efficiency gains from removing quotas in terms of observed profits and the inverse of the semi-elasticities matrix H. The intuition for (8) comes from the fact that profits of unconstrained firms are zero. Thus, we can express the efficiency gains from removing quotas in terms of their initial profits and the rate at which profits change as the quotas are relaxed (described by H).

The expressions in Proposition 7 can be used to estimate the efficiency gains from relaxing all or any subset of quotas. Corollary 1 specializes the expressions from Proposition 7 to the case of removing a single quota.

Corollary 1 (Efficiency gains from removing a single quota)**.** *Let Π_i be the profits of producer i, and let $d\log y_{i^*} = \log y_{i^*} - \log y_i^{eff}$ be the log-difference between the quota on i and the level of i's output that would obtain without a quota, holding quotas on all other producers fixed. The efficiency gains from removing the quota on producer i up to the second order in $d\log y_{i^*}$ can be estimated using any of the three following expressions:*

$$\Delta\log Y \approx -\frac{1}{2}\Pi_i d\log y_{i^*}. \qquad \text{(Option 1)}$$

$$\Delta\log Y \approx -\frac{1}{2}\frac{\partial\Pi_i}{\partial\log y_{i^*}}(d\log y_{i^*})^2. \qquad \text{(Option 2)}$$

$$\Delta \log Y \approx \frac{1}{2}\Pi_i \left[-\frac{d \log \Pi_i}{d \log y_{i^*}} \right]^{-1} \qquad \text{(Option 3)}$$

The expressions labelled Options 1–3 in Corollary 1 correspond to the equations (6)–(8) in Proposition 7. Note that in the final expression, labeled Option 3, we instead express the efficiency gains from removing a quota in terms of the elasticity of profits with respect to the quota (rather than the semi-elasticity). The efficiency gain is inversely related to the elasticity of profits with respect to the quota because, fixing the level of initial profits, if profits fall quickly as the quota is relaxed, a small change in the quota level is required to take the economy to the unconstrained point. Conversely, if profits fall slowly as the quota is relaxed, the distance to the unconstrained point is large, since it will take a large change in quantity to restore profits to zero.

The elasticity $d \log \Pi_i / d \log y_{i^*}$ can also be useful to differentiate empirically between situations where the quota on a producer is close to or far from its unconstrained level of production. If the quota is close to the unconstrained level, the elasticity $d \log \Pi_i / d \log y_{i^*}$ must be negative, since profits must fall to zero as the level of the quota rises to the point where it is no longer binding. Hence, if the elasticity $d \log \Pi_i / d \log y_{i^*}$ at an initial equilibrium is positive—i.e., an increase in the output allowed from a sector leads to an increase in the sector's profits—then the economy must be far from the efficient frontier. In this case, the assumption that the quantity distortion is small is violated, and the expressions in Corollary 1 cease to be a reasonable approximation for the efficiency gains.

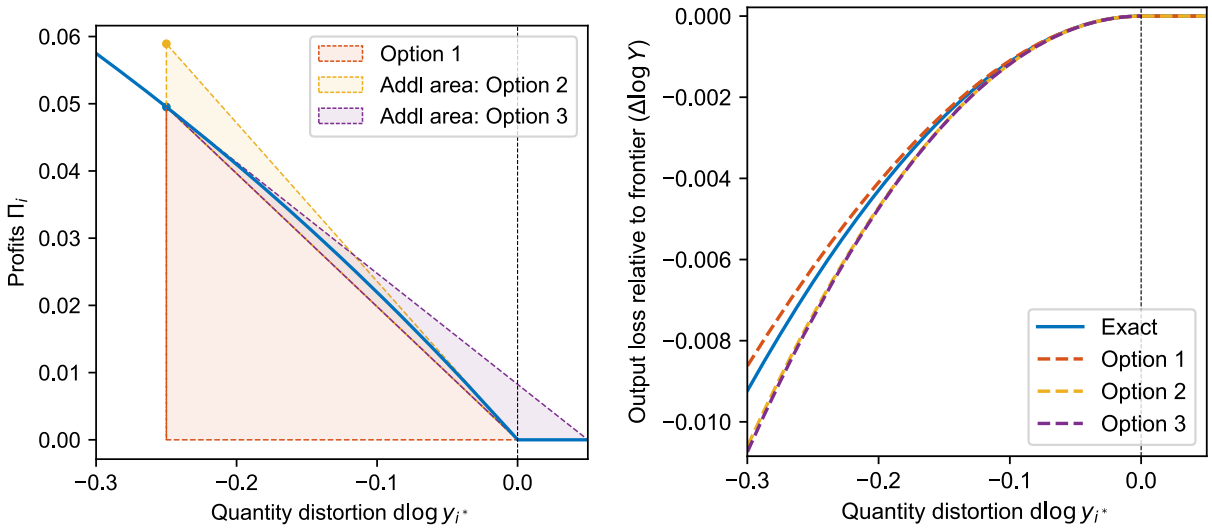
Example 9 (Round-About Economy)**.** We illustrate the effects of removing a single quota in a round-about economy. There is a single firm i that produces using labor and its own goods. The elasticity of substitution between labor and its own goods in production is θ. A quota limits the amount of the round-about firm's output that can be used as an input in production. We apply each of our three expressions for the distance to the frontier in turn.

First, Equation (6) shows that we can estimate the distance to the frontier using the profits of the constrained producer and the size of the distortion,

$$\Delta \log Y \approx -\frac{1}{2}\Pi_i d \log y_{i^*}. \qquad \text{(Option 1)}$$

In Figure 4, for a given quantity distortion $d \log y_{i^*}$, the estimated distance to the frontier is given by multiplying the quantity distortion by the resulting profits Π_i and one-half. This formula approximates the area under the profit function and thus the output gains from moving to the efficient frontier.

Figure 4: Distance to the frontier in a round-about economy.



(a) Profits.

(b) Distance to the frontier.

Second, Equation (7) replaces the level of profits, Π_i, with the semi-elasticity of profits with respect to the quota times the size of the distortion,

$$\Delta \log Y \approx -\frac{1}{2}\frac{d\Pi_i}{d\log y_{i^*}}(d\log y_{i^*})^2 = \frac{1}{2\theta}\frac{\lambda_i - 1}{\lambda_i}(d\log y_{i^*})^2. \qquad \text{(Option 2)}$$

The second equality expresses the semi-elasticity of profits with respect to the quota in terms of the sales of the round-about firm, λ_i, and the elasticity of substitution between labor and the round-about input in firm 1's production function, θ_1. In Figure 4, this approximation for the distance to the frontier corresponds to estimating profits by extrapolating the profit function out from the efficient point where $d\log y_{i^*} = 0$, and then multiplying those estimated profits by the size of the distortion $d\log y_{i^*}$ and one-half.

Third, Equation (8) estimates the size of the distortion, $d\log y_{i^*}$, by estimating the local elasticity of profits to quota changes around the initial, distorted allocation,

$$\Delta \log Y \approx \frac{1}{2}\Pi_i \left[-\frac{d\log \Pi_i}{d\log y_{i^*}}\right]^{-1}. \qquad \text{(Option 3)}$$

Starting with a given distortion $d\log y_{i^*}$, this approximation uses the level of profits Π_i and estimates the size of the distortion $d\log y_{i^*}$ by extrapolating the profit function out from the inefficient point. As shown in the right panel of Figure 4, this expression, as well as the other two alternatives, closely approximates the true distance to the frontier even

as the quantity distortion becomes large.

Example 10 (Horizontal Economy with Multiple Quotas)**.** Consider again the horizontal economy with quotas y_{1^*} and y_{2^*} from Example 8. Applying Proposition 7 shows the efficiency gains from relaxing both quotas y_{1^*} and y_{2^*} are

$$\Delta \log Y \approx -\frac{1}{2}\left(\Pi_1^2 H_{11}^{-1} + \Pi_1^2 H_{22}^{-1}\right) - \Pi_1 \Pi_2 H_{12}^{-1}.$$

The final term, $-\Pi_1 \Pi_2 H_{12}^{-1}$, describes the additional efficiency gain that results from relaxing both quotas together compared to the sum of the efficiency gains realized from relaxing each quota individually. If H_{12}^{-1} is positive, the gains from relaxing one quota partially offset the gains from relaxing the other. On the other hand, if H_{12}^{-1} is negative, relaxing each quota amplifies the additional efficiency gains associated with the other.

Since the matrix H is negative definite at the efficient point, the sign of H_{12}^{-1} near efficiency is given by the sign of $-H_{12} = -\partial \Pi_1 / \partial \log y_{2^*}$. Since profits at the efficient point are zero, it must always be the case that $H_{12} = \partial \Pi_1 / \partial \log y_{2^*} \leq 0$. Thus, around the efficient point, relaxing the quota on firm 2 always weakly decreases the profits of firm 1, H_{12}^{-1} is weakly positive, and the gains from relaxing the two quotas must always (weakly) offset each other.

6 Empirical Applications

We demonstrate how to apply our results in several empirical examples. The first two empirical examples, which consider the cap on H-1B visas and zoning restrictions on single-family housing, illustrate how to apply our results on the first-order effects of quota changes from Section 3. The following three examples, on Argentina's capital controls, U.S. quotas on Chinese textile and clothing exports, and taxicab medallions in New York City, each illustrate various results on nonlinearities and the distance to the frontier from Sections 4 and 5.

6.1 H-1B Visa Quota

The H-1B visa allows U.S. firms to employ high-skill foreign workers. Since the mid-2000s, the total number of visas issued has been capped at 85,000, with 20,000 of the slots reserved for immigrants holding a master's or higher degree from a U.S. university. We can use our results to estimate the world efficiency gains that would result from relaxing the cap on H-1B visa quotas.

Our measure of the profits that accrue to winners of the H-1B visa lottery comes from Clemens (2013), who compares earnings of winners and losers of the 2007 H-1B lottery within a pool of Indian software workers employed at the same firm. In 2007, the U.S. government received more applications than needed to fill the H-1B quota within the first two days of the application window and chose which H-1B visa applications to process by random lottery. Earnings for workers whose applications were processed—those who won the lottery—were $12,641 higher two years after the lottery than their colleagues who lost the lottery.

If we assume that software workers are paid their marginal product, then the first-order efficiency gains from expanding the H-1B cap can be computed from this statistic alone. We apply Proposition 3 to get

$$d \log Y = \Pi_i d \log y_{i^*} \approx \frac{\Pi_i}{y_{i^*}} dy_{i^*}.$$

That is, the efficiency gain in dollars from increasing the H-1B cap by one slot is equal to the per-person rents of visa holders today. This means that for example, doubling the number of available visas in 2007 would have increased world output by $1.07B.

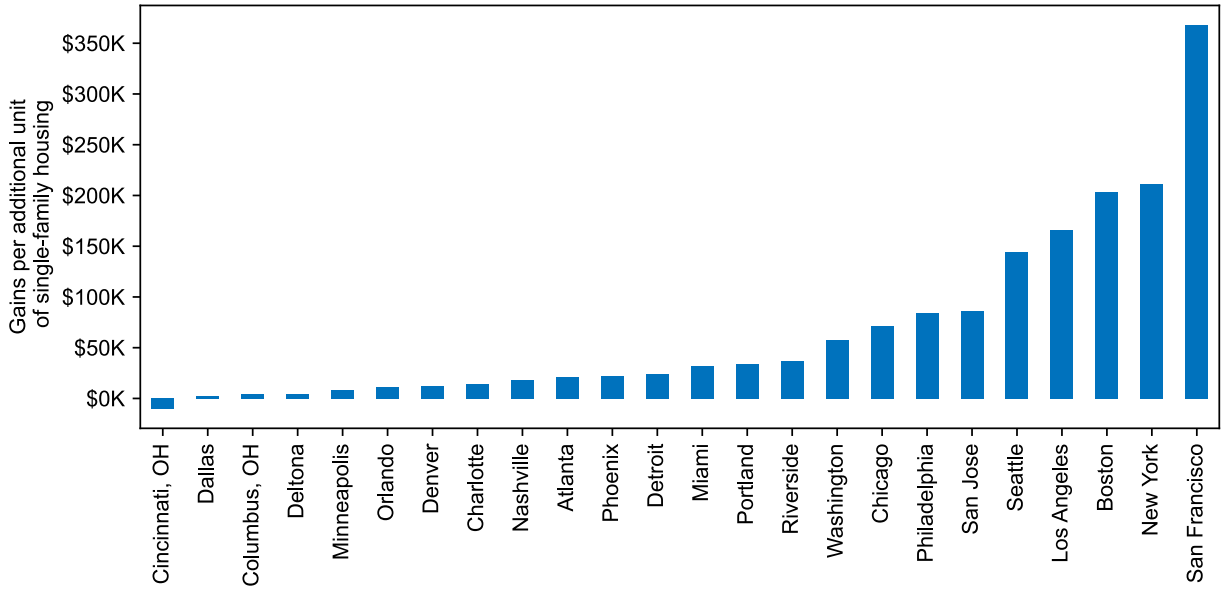
Note that this figure reflects efficiency gains in *world* output from increasing the number of H-1B visas. It does not include reallocations in output from the rest of the world to the U.S., e.g. from moving workers to the U.S. from other countries. Assuming all other distortions take the form of quotas and are held fixed, the additional increase in U.S. output from moving workers to the U.S. is equal to the workers' earnings minus the rents they receive, $85{,}000 \times \$31{,}194 \approx \$2.7\text{B}$ (using earnings of lottery losers from Clemens 2013).

6.2 Zoning Restrictions on Single-Family Housing

Next, consider the potential efficiency gains from relaxing zoning restrictions on single-family housing across U.S. cities. To estimate the rents that accrue to zoning restrictions, we use data on "zoning taxes" for 24 metropolitan statistical areas (MSAs) from Gyourko and Krimmel (2021). They measure these zoning taxes by comparing land prices for vacant parcels purchased to build new single-family housing units—which include the rights to supply single-family housing—with land prices on nearby parcels that have existing single-family homes. This comparison isolates the value of permits to build a new single-family housing unit from the value of the land itself.

Figure 5 shows the estimated gains associated with relaxing zoning restrictions to

Figure 5: Gains from expanding the supply of single-family housing across U.S. cities.



increase the supply of single-family housing in each MSA.[8] Supplying an additional unit of single-family housing is associated with efficiency gains of over $350,000 in San Francisco, and over $150,000 in other coastal cities like New York, Boston, and Los Angeles.

Policymakers often state housing policies in terms of the number of permits they plan to make available, as these permits directly control the supply of housing in zoning-constrained cities.[9] Modeling zoning restrictions as quantity distortions allows one to map these proposals to expand the supply of housing permits directly into efficiency gains. Moreover, modeling zoning restrictions as quotas has the advantage of requiring less information than modeling them as wedge distortions. Using the wedge approach, we would need to estimate the reduction in zoning wedges necessary to achieve a target increase in housing, which depends on underlying elasticities of supply and demand for housing across U.S. cities. In contrast, Proposition 3 allows us to directly use proposed quantity changes without having to map from quantities to wedges and back.

[8]Gyourko and Krimmel (2021) observe several vacant parcel sales in each MSA. To estimate zoning rents per unit of single-family housing in each MSA, we use the median of estimated zoning taxes per quarter acre in each MSA and divide this estimate by the median acreage of single-family homes in the MSA.

[9]For example, California state mandates require that San Francisco approve the creation of 82,000 new housing units by 2031. See `https://www.sfchronicle.com/projects/2023/san-francisco-housing/`.

6.3 Argentina's Capital Controls

We use Proposition 7 to estimate the distance to the frontier in the context of restrictions on capital outflows imposed by Argentina. On September 1, 2019, the Argentine government reimposed capital controls following a four-year period with no restrictions on capital flows. The restrictions initially limited US dollar purchases by individuals to $10,000 per month and imposed tighter controls on corporate access to foreign exchange. Following this imposition of capital controls, capital outflows fell from an average of $7.2B per month in the free market period to under $1.5B.

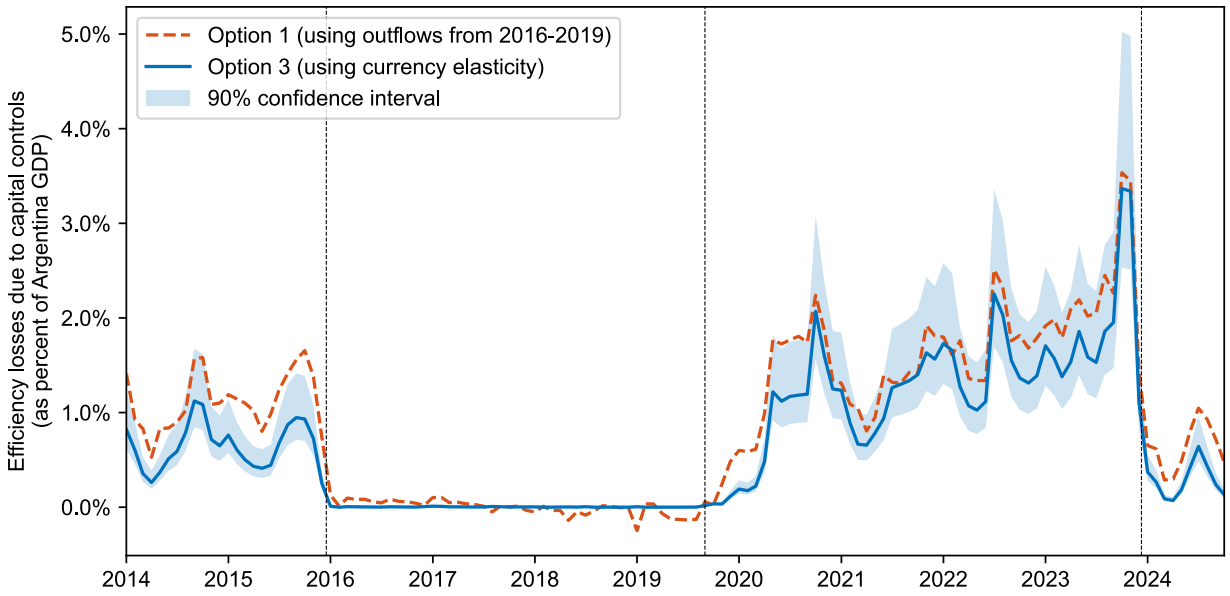
We use two approaches to estimate the efficiency losses due to these quotas on capital outflows. The first approach applies Option 1, which expresses the distance to the frontier in terms of the profits accruing to quota holders and the size of the distortion. In the context of Argentina, transactions that are permitted under the capital outflow restrictions exchange Argentine pesos for dollars at the official exchange rate, which grants pesos a substantial premium relative to black market exchange rates. We can measure the profits that accrue to quota holders who are permitted to make transactions using the gap between black market and official exchange rates, $\Pi_i = (\log e/\bar{e})\, y_{i^*}$, where e and $\bar{e}$ are the black market and official Argentina peso–USD exchange rates and y_{i^*} is the allowed quantity of capital outflows. Thus, Proposition 7 Equation (6) becomes

$$\Delta \log Y \approx -\frac{1}{2}\Pi_i d \log y_{i^*} \approx -\frac{1}{2}(\log e/\bar{e})\, dy_{i^*}.$$

The dashed line in Figure 6 plots the distance to the frontier estimated using Option 1. We use the most popular black market exchange rate, known as the "Dólar Blue," to measure the profits accruing to quota holders with the license to exchange pesos at the official exchange rate. We measure the size of the distortion dy_{i^*} as the difference between the (restricted) level of capital outflows and the average level of outflows during the period without capital controls from January 2016 to August 2019. Since the reinstatement of capital controls in September 2019, the estimated efficiency losses due to capital controls average 1.4 percent of Argentina's GDP and reach a high of 3.5 percent of GDP just before the devaluation of the peso in late 2023.

A disadvantage of this first approach is the strict assumption that the efficient level of capital outflows during the period with capital controls is equal to the observed level of outflows during the period without controls. Our second approach instead uses Option 3 to back out the size of the distortion using the level of profits and the responsiveness of profits to outflows. For these restrictions on capital outflows, we can measure the responsiveness of profits to outflows by estimating how allowing for additional outflows

Figure 6: Efficiency losses due to Argentina's capital controls.



Note: The three vertical dashed lines correspond to the end of capital controls on December 17, 2015, the reinstatement of capital controls on September 1, 2019, and the devaluation of the peso by the Milei government on December 10, 2023. The wedge between market and official Argentine exchange rates is calculated using the Dólar Blue and official exchange rates from Refinitiv. Option 1 calculates the size of the distortion as the difference in monthly capital outflows relative to the average from Jan 2016 to Sep 2019, using data from the Central Bank of Argentina (BCRA). Option 3 applies the currency elasticity and standard errors estimate by Adler et al. (2019).

would change the official exchange rate and thus shrink the gap between the black market and official exchange rates.

A common statistic used to summarize the responsiveness of exchange rates to outflows is the depreciation in nominal exchange rates caused by purchases of foreign currency equal to one percent of GDP (Blanchard et al. 2015; Adler et al. 2019). Denoting the *currency elasticity* of nominal exchange rates to outflows as a share of GDP by θ, we can express the distance to the efficient level of capital outflows as

$$dy_{i^*} = \frac{1}{\theta} \text{GDP} \left(\log e/\bar{e}\right).$$

Lower values of θ imply a greater size of distortion, since more capital outflows would be required to close the gap between the black market exchange rate e and the official exchange rate $\bar{e}$.

Combining this expression with the previous, we can express the efficiency losses due

to capital controls as a share of Argentina's GDP in terms of the currency elasticity θ and the gap between market and official exchange rates,

$$\frac{\Delta Y}{\text{GDP}} \approx -\frac{1}{2}\frac{1}{\theta}\left(\log e/\bar{e}\right)^2.$$

The distance to the frontier is greater when the current elasticity θ is low. The distance to the frontier also scales quadratically in the gap between black market and official exchange rates, because a higher gap implies both higher profits per dollar of capital flow and implies a greater quantity distortion relative to the frontier.

The solid line in Figure 6 plots the distance to the efficient frontier over time using this second approach, applying estimates of the currency elasticity θ from Adler et al. (2019).[10] The efficiency losses due to capital controls estimated using this approach line up closely with the estimates of the distance to the frontier from Option 1. The estimates again indicate substantial efficiency losses, for example averaging 1.9 percent of Argentina's GDP over 2023. The estimates also indicate that changes since late 2023 have substantially lowered the distance to the frontier. A sharp devaluation of the peso on December 13, 2023, instituted as part of Milei's shock therapy economic plan, lowered the efficiency losses to below 0.5 percent of GDP. Growing investor confidence in late 2024 also narrowed the gap between the black market and official exchange rates, despite the fact that permitted capital outflows have remained low, narrowing the distance to the frontier to under 0.2 percent of GDP in October 2024.

6.4 U.S. Quotas on Chinese Textile & Clothing Exports

We illustrate how the interaction of multiple quotas affects efficiency gains using the phase-out of textile and clothing quotas under the World Trade Organization (WTO) Agreement on Textile and Clothing (ATC). From 1975 to 1994, the Multi-Fiber Agreement (MFA) had imposed quotas on exports of textiles and clothing from developing countries to the US and the EU. These quotas were particularly binding on China—whose textile and clothing exports to the US rose dramatically when these quotas were relaxed—as well as other Asian exporters such as Bangladesh, India, Singapore, and Hong Kong (Dean

[10] Adler et al. (2019) estimate that outflows equal to one percent of GDP lead to 1.7–2.0 percent depreciation in nominal exchange rates. For Argentina, these estimates imply that $1B of outflows in 2023 results in a depreciation in the Argentine peso by 0.26%. These estimates align with previous work: for example, using exogenous global capital flow shocks, Blanchard et al. (2015) estimate that outflows equal to one percent of GDP lead to a 1.5% depreciation in nominal exchange rates. Estimates of the impact of order flows on currency markets are also quantitatively similar. For example, Evans and Lyons (2002) find that $1B of net purchases in 1996 leads to an 0.54% appreciation (or, converting to 2023 dollars, $1B in 2023 USD outflows leads to a currency depreciation of 0.30%).

1990). As part of the WTO's Uruguay Round, the Agreement on Textile and Clothing (ATC) introduced a plan for phasing out these quotas over the period from 1995 to 2005.

The removal of quotas on textile and clothing goods in phases over this period allows us to study the interactions between sets of quotas. We focus in particular on quotas on China, and on the interaction between the quotas that were lifted as part of Phase III of the ATC in 2002 and quotas lifted in Phase IV of the ATC in 2005.[11] Goods with quotas lifted in Phase III included knit fabrics, gloves, dressing gowns, brassieres, and textile luggage products; while a broader set of quotas on silk, wool, and cotton textiles, carpets, and most apparel categories were not lifted until Phase IV in 2005.

We estimate the effects of the Phase III and Phase IV quota removals on exports using the specification,

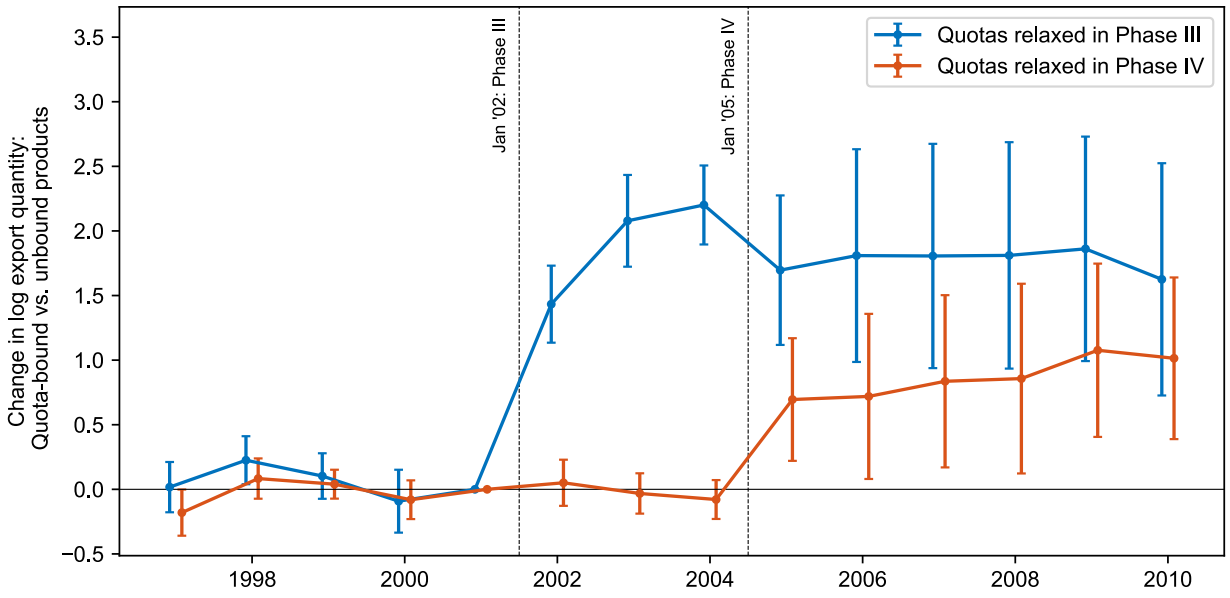
$$\begin{aligned}\log y_{ict} = {} & \beta_t^{\text{Phase III}} \Big(\text{Binding}_c \times 1\{c \text{ quota relaxed in Phase III}\} \times 1\{\text{year} = t\}\Big) \\ & + \beta_t^{\text{Phase IV}} \Big(\text{Binding}_c \times 1\{c \text{ quota relaxed in Phase IV}\} \times 1\{\text{year} = t\}\Big) + \alpha_t + \delta_i + \varepsilon_{ict}, \quad (9)\end{aligned}$$

where y_{ict} is the quantity of exports of HS-10 code i in category c from China to the US in year t, Binding_c indicates whether the quota on category c was initially binding, and α_t and δ_i are year and HS-10 code fixed effects. Note that specification (9) measures changes in export quantities for goods with initially binding quotas relative to other goods also included in the ATC whose quotas were non-binding. Following Brambilla et al. (2010), we define a quota as binding if the fill rate (i.e., realized exports as a percent of the quota allowance) exceeds 90 percent. We estimate (9) using data on Chinese exports to the US at the HS-10 level from the Office of Textiles and Apparel (OTEXA) and data on quota fillrates from the US MFA/ATC database created by Brambilla et al. (2010).

Figure 7 plots the estimated coefficients for $\beta_t^{\text{Phase III}}$ and $\beta_t^{\text{Phase IV}}$ from specification (9). Phase III of the ATC in 2002 led to a large increase in exports for products whose quotas expired in 2002. Exports for HS-10 codes in the Phase III group with initially binding quotas rose by more than 180 log points from 2002–2004 relative to products with non-binding quotas. The final Phase IV of the ATC in 2005 led to a small decline in exports for Phase III group products relative to 2002–2004, and an 80 log point rise in exports for HS-10 codes in the Phase IV group.

[11] Although the ATC officially required quotas to be removed in four phases from 1995 to 2005, the structure of the agreement allowed the US (and the EU) to defer the removal of most binding quotas until the final two phases of the agreement. During Phase I (1995) and Phase II (1998), the US strategically liberalized non-binding quotas or low-restriction categories; the real impact of the ATC materialized in Phase III (2002) and Phase IV (2005), when the US began lifting quotas that had been actively constraining trade (Chiron 2004).

Figure 7: Differential changes in export quantity for products with initially binding quotas.



Note: The blue and red lines plot estimates for $\beta_t^{\text{Phase III}}$ and $\beta_t^{\text{Phase IV}}$, respectively, from specification (9). The sample includes 14,975 observations across 1,931 HS-10 codes. Standard errors are two-way clustered by category and year. Error bars indicate 95 percent confidence intervals.

We combine these estimates with data on quota license prices to estimate the matrix of semi-elasticities of profits to quota changes.[12] We measure the initial aggregate profits of quota holders for Phase III and Phase IV products by multiplying quota license prices in 2001 by the quantity of exports in those product categories in 2001. Assuming that profits for Phase III and Phase IV group products go to zero when quotas are relaxed, we can then solve for the matrix H by solving the following system of equations:

$$\begin{aligned}
\Pi_{\text{Phase III}} &= \beta_{\text{III}}^{\text{Phase III}} H_{11}, \\
\Pi_{\text{Phase III}} &= \beta_{\text{IV}}^{\text{Phase III}} H_{11} + \beta_{\text{IV}}^{\text{Phase IV}} H_{12}, \\
\Pi_{\text{Phase IV}} &= \beta_{\text{IV}}^{\text{Phase III}} H_{12} + \beta_{\text{IV}}^{\text{Phase IV}} H_{22},
\end{aligned}$$

where $\beta_{\text{III}}^{\text{Phase III}}$ is the effect of relaxing the Phase III quotas on Phase III products' exports, $\beta_{\text{IV}}^{\text{Phase III}}$ is the effect of relaxing the Phase IV quotas on Phase III products' exports, and $\beta_{\text{IV}}^{\text{Phase IV}}$ is the effect of relaxing the Phase IV quotas on Phase IV products' exports. First, the increase in export quantities for Phase III products from 2002–2004 identifies the semi-elasticity of profits for Phase III products to their quotas, holding the Phase IV quotas

[12] We are grateful to Amit Khandelwal and Judith Dean for sharing data on these quota prices, which were originally scraped from chinaquota.com.

Table 1: Gains from relaxing textile/clothing quotas on Chinese exports to the US.

Intervention	Estimated efficiency gains (2001 USD billions)
(A) Relaxing Phase III quotas only	\$40
(B) Relaxing Phase IV quotas only	\$158
(C) Relaxing both Phase III and IV quotas	\$185
Difference: C − (A + B)	\$13

fixed.[13] Second, the change in exports of both Phase III products after 2005 allows us to estimate the cross-product elasticity H_{12}. Finally, since the symmetry of H guarantees $H_{12} = H_{21}$, we can estimate the semi-elasticity of profits for Phase IV products to their quotas, H_{22}.[14] Solving this system of equation yields

$$\Pi = \begin{bmatrix} \Pi_{\text{Phase III}} \\ \Pi_{\text{Phase IV}} \end{bmatrix} = \begin{bmatrix} \$38\text{B} \\ \$394\text{B} \end{bmatrix}, \qquad H = \left[d\Pi_i / d \log y_j \right] = \begin{bmatrix} -17.9 & -7.6 \\ -7.6 & -495.6 \end{bmatrix}$$

Note that the off-diagonal entry H_{12} is negative. This negative cross-term is identified by the decline in export quantities for Phase III products when Phase IV quotas were lifted in 2005. As discussed in the analytic example above, $H_{12} < 0$ implies $H_{12}^{-1} > 0$ and that the gains from relaxing both the Phase III and Phase IV quotas together are smaller than the sum of the gains from relaxing each subset of quotas individually.

This prediction is borne out in Table 1, which estimates the efficiency gains from either relaxing each set of quotas individually to the gains or relaxing all the MFA quotas together by applying Proposition 7 Equation (8). Starting from the quota levels in 2001, we estimate that relaxing either the Phase III or Phase IV quotas would have increased efficiency by \$40 and \$158 billion, respectively. However, relaxing all quotas increases together efficiency by \$185 billion, or \$13 billion less than the sum of the gains from relaxing each set of quotas in isolation.

[13] While the phases of the ATC technically required changes in Phase IV products' quota levels even before the quotas were completely relaxed in 2005, we assume that Phase IV quotas were held fixed since our estimates of $\beta_t^{\text{Phase IV}}$ for $t \in \{2002, 2003, 2004\}$ are not significantly different from zero.

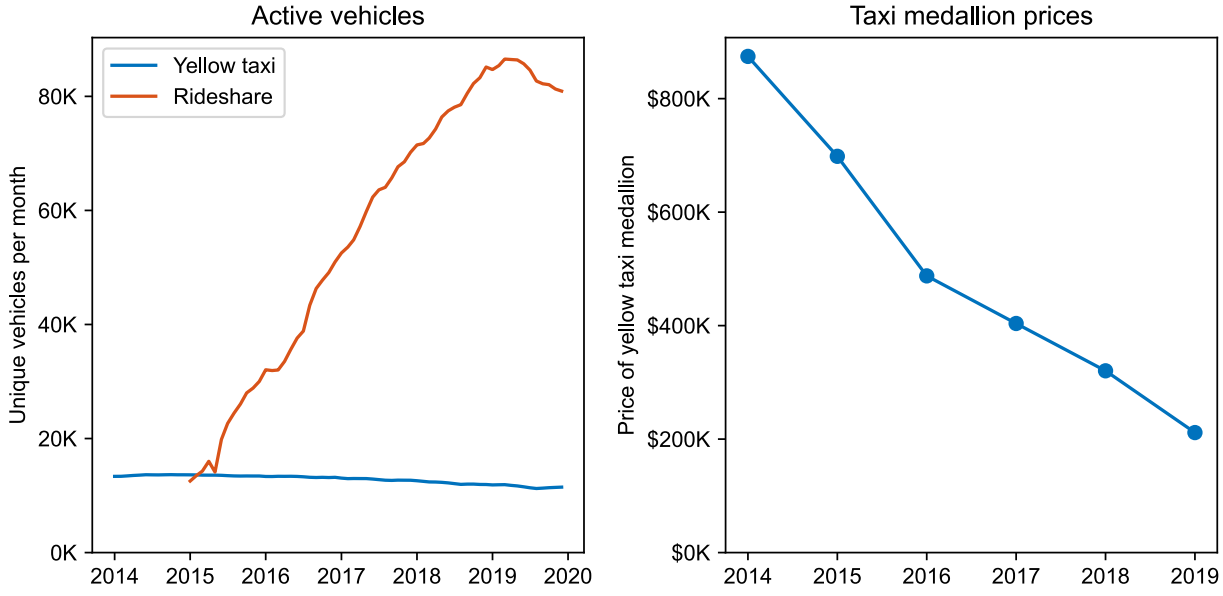
[14] US textile and clothing industry groups lobbied for new quotas on a subset of categories after 2005, though the new quotas were in most cases substantially higher than the expiring ATC quotas. We find similar quantitative results if we exclude products that had quotas imposed after 2005 from our estimation.

6.5 Taxicab Medallions in New York City

We apply our results to estimate the welfare losses due to taxicab medallion quotas in New York. Taxicab medallions are required to operate a taxi; the city of New York created the taxicab medallion system in 1937 to restrict the total supply of taxicabs. We exploit the growth of rideshare apps such as Uber and Lyft in New York to estimate the efficiency gains from relaxing these restrictions on the supply of taxis.

The first panel of Figure 8 shows how the number of taxi and rideshare vehicles in New York from 2014 to 2019. The number of unique taxis active each month has stayed around 13,000, just under the total 13,587 taxi medallions available from the New York Taxi and Limousine Commission. The number of rideshare vehicles, on the other hand, grew nearly sevenfold from about 12,500 in January 2015 to over 85,000 by mid-2019. During this time, the transaction prices of taxi medallions also fell dramatically, from nearly $1 million dollars at its peak in 2014 to $200,000 in 2019.[15]

Figure 8: Changes in New York taxi market from 2014–2019.

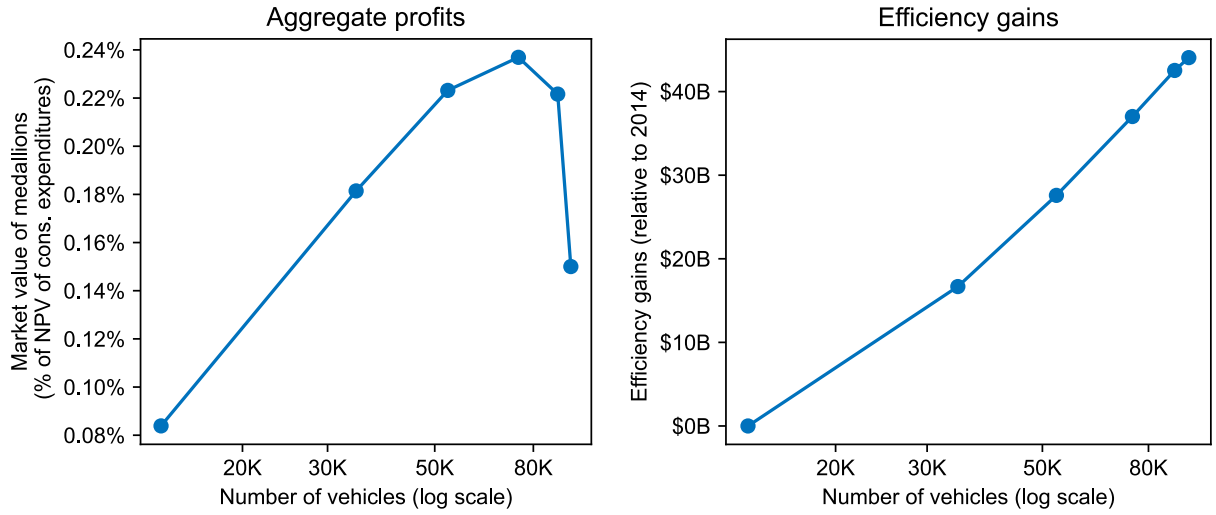


Note: Monthly unique vehicles are from aggregated reports from the NYC Taxi and Limousine Commission. Taxi medallion prices are annual averages of prices for medallion transfers, from the NYC Taxi and Limousine Commission.

We use how taxi medallion prices fall with the entry of rideshare vehicles to estimate

[15]Similar trends unfolded in other U.S. cities when rideshare apps entered the market. For example, medallion prices in both Boston and Chicago dropped 30–40 percent from 2015 to 2016. See `https://www.foxnews.com/opinion/are-taxi-medallions-too-big-to-fail-too`.

Figure 9: Profits and efficiency gains in the New York taxi market from 2014–2019.



Note: Aggregate profits are medallion transaction prices times number of active vehicles. Profits are shown as a share of the NPV of consumer expenditures, calculated using BLS Consumer Expenditure Surveys Northeast MSA statistics with a 4% discount rate. Efficiency gains are calculated by cumulating (10).

the output gains from relaxing the quota on the number of taxis in New York. For this exercise, we make two assumptions. First, we assume that ride-sharing services are a perfect substitute to taxis, and hence the introduction of ride-sharing services is equivalent to relaxing the quota on the number of vehicles in the market. Second, we assume that taxi medallion prices reflect the discounted value of all future profits accruing to owners of vehicles that are approved to provide rides in New York.[16]

The left panel of Figure 9 shows aggregate profits accruing to taxis and rideshare vehicles as the number of vehicles increased from 2014 to 2019. The number of vehicles in the market was initially so low that initial increases in the number of vehicles in fact increased the aggregate profits earned by these vehicles. Since initially $d\Pi_i/d\log y_{i^*} > 0$, the market was in the region where output is convex with respect to quota changes (as seen from Proposition 5). Moreover, the fact that aggregate profits rose as the quota was relaxed means that the initial number of medallions was below the level that a monopolist would have chosen.

Following Proposition 5, we can approximate the efficiency gains from relaxing the

[16]We find similar results if we instead calculate taxicab drivers' excess profits using the change in taxis' revenues as Uber and Lyft entered the market. From 2014 to 2019, revenues per taxi fell by about $40,000 annually, while the change in taxi medallion prices over this period corresponds to a decline in annual profits per taxi of approximately $37,000. The advantage of using medallion prices is that they isolate changes in profits expected to accrue to medallion owners from other changes in costs that affect revenues.

Table 2: Estimated efficiency gains from relaxing capacity constraint on New York taxis.

	Change from 2014–2019	Distance to frontier
Output gains	\$44.1B	\$1.8B
Gains per New York MSA household	\$6,029	\$246
% of NPV of transportation expenditures (incl. vehicles/gas)	2.61%	0.11%

Note: New York MSA consumer units and transportation expenditures are from the BLS Consumer Expenditure Surveys 2018–2019 northeast MSA statistics. The net present value of transportation expenditures is calculating using annual transportation expenditures in 2018–2019 and a 4% discount rate.

medallion quota to a second order in each year using

$$\Delta \log Y_t \approx \left(\Pi_{it} + \frac{1}{2} d\Pi_{it}\right) d \log y_{i^* t}. \tag{10}$$

As shown in Figure 9, these gains are largest in 2014 and 2015 as ride-share vehicles first enter the market, and by 2019 cumulate to nearly \$44B in efficiency gains. The first column of Table 2 shows that these gains translate into \$6,029 per household in the New York City metro area, or 2.6% of the present value of current and future household transportation expenditures.[17]

Of course, even in 2019, the market is not efficient, since the supply of vehicles is determined by the number of medallions and by profit-maximizing ride-share companies. We use Proposition 7 Equation (8) to estimate the distance to the frontier in 2019, using the level of aggregate profits in 2019 and the elasticity of aggregate profits to changes in the number of vehicles from 2018 to 2019. The second column of Table 2 shows that the remaining distance to the frontier is small compared to the efficiency gains achieved from 2014 to 2019. In particular, increasing the number of vehicles toward the efficient level would only add a further \$246 in gains per household in the New York MSA.

7 Extensions

In this section, we describe extensions of our framework that are developed in the Online Appendix.

[17] Our estimates correspond to an annual gain around \$1.8B (assuming a 4% discount rate). These gains are a similar order of magnitude to the consumer surplus gains estimated by Cohen et al. (2016), who use estimates of consumer price elasticity along the demand curve for Uber rides to calculate an annual consumer surplus of \$2.9B from Uber across four cities (New York, Los Angeles, Chicago, and San Francisco).

Analytic expressions for a CES economy. Our results characterize the distance to the frontier and the nonlinear effects of quotas in terms of semi-elasticities of profits with respect to quotas. In Appendix A, we provide explicit formulas for this inverse demand system in terms of the input-output matrix and microeconomic elasticities of substitution. For these expressions, we focus on general input-output economies in which all nodes have constant elasticity of substitution production functions.

Rent-seeking. In Appendix B, we extend the framework to allow for rent-seeking, in which quotas are enforced through permits and households expend productive resources to acquire permits. Whether a quota diverts households to expend their labor on rent-seeking rather than production depends on whether the costs of permits, set exogenously by the government, are set below a certain threshold. When there is free entry into rent-seeking, as in Krueger (1974), the comparative statics of output with respect to quota changes include an additional term that depends on how the quota change affects labor income and quota profits earned in excess of permit costs. This rent-seeking effect can result in first-order losses associated with quota changes even starting at the efficient allocation.

8 Conclusion

This paper develops a framework for analyzing economies with quotas and other quantity-based distortions. These economies are constrained efficient, allowing us to develop non-parametric results for the effects of microeconomic shocks and the misallocation costs due to quotas that rely on a small set of sufficient statistics.

We propose that these results can be used to evaluate policy counterfactuals and to characterize the social costs of quota distortions in many settings. Our sufficient statistics approach allows one to estimate counterfactuals without information on the input-output matrix and microeconomic elasticities of substitution that are generally needed to compute the effects of shocks in economies with wedges. The empirical examples we develop in this paper—applied to H-1B visas, zoning restrictions, capital controls, import quotas, and taxicab medallions—illustrate how one can measure the sufficient statistics necessary to apply our results.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

Adler, G., N. Lisack, and R. C. Mano (2019). Unveiling the effects of foreign exchange intervention: A panel approach. *Emerging Markets Review* (100620).

Anderson, J. E. (1985). The relative inefficiency of quotas: The cheese case. *American Economic Review 75*(1), 178–190.

Atalay, E. (2017). How important are sectoral shocks? *American Economic Journal: Macroeconomics 9*(4), 254–280.

Baqaee, D. R. (2018). Cascading failures in production networks. *Econometrica 86*(5), 1819–1838.

Baqaee, D. R. and A. Burstein (2023). Welfare and output with income effects and taste shocks. *The Quarterly Journal of Economics 138*(2), 769–834.

Baqaee, D. R. and E. Farhi (2019). The macroeconomic impact of microeconomic shocks: beyond hulten's theorem. *Econometrica 87*(4), 1155–1203.

Baqaee, D. R. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics 135*(1), 105–163.

Baqaee, D. R. and E. Rubbo (2023). Micro propagation and macro aggregation. *Annual Review of Economics 15*(1), 91–123.

Basu, S. and J. G. Fernald (2002). Aggregate productivity and aggregate technology. *European Economic Review 46*(6), 963–991.

Bau, N. and A. Matray (2023). Misallocation and capital market integration: Evidence from india. *Econometrica 91*(1), 67–106.

Bhagwati, J. (1965). *Trade, growth and the balance of payments*, Chapter On the equivalence between tariffs and quotas, pp. 53–67. Rand McNally.

Bigio, S. and J. La'O (2020). Distortions in production networks. *The Quarterly Journal of Economics 135*(4), 2187–2253.

Blanchard, O., G. Adler, and I. de Carvalho Filho (2015). Can foreign exchange intervention stem exchange rate pressures from global capital flow shocks. IMF Working Paper.

Boorstein, R. and R. C. Feenstra (1991). *International Trade and Trade Policy*, Chapter Quality upgrading and its welfare cost in U.S. steel imports, 1969–74, pp. 167–186. The MIT Press.

Brambilla, I., A. K. Khandelwal, and P. K. Schott (2010). *China's Growing Role in World Trade*, Chapter China's Experience under the Multi-Fiber Arrangement (MFA) and the Agreement on Textiles and Clothing (ATC), pp. 345–387. University of Chicago Press.

Buera, F. J. and N. Trachter (2024). Sectoral development multipliers. Technical Report 32230, National Bureau of Economic Research.

Carvalho, V. M. and A. Tahbaz-Salehi (2019). Production networks: A primer. *Annual Review of Economics 11*(1), 635–663.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007). Business cycle accounting. *Econometrica 75*(3), 781–836.

Chiron, C. (2004). Influences of quotas, tariffs, and bilateral trade agreement on post 2005 apparel trade. Technical report, Harvard Center for Textile and Apparel Research.

Clemens, M. A. (2013). Why do programmers earn more in houston than hyderabad? evidence from randomized processing of us visas. *American Economic Review 103*(3), 198–202.

Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe (2016). Using big data to estimate consumer surplus: The case of uber. Technical Report 22627, National Bureau of Economic Research.

Dasgupta, P. and J. E. Stiglitz (1977). Tariffs vs. quotas as revenue raising devices under uncertainty. *American Economic Review 67*(5), 975–981.

De Loecker, J., P. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Dean, J. M. (1990). The effects of the US MFA on small exporters. *The Review of Economics and Statistics 72*(1), 63–69.

Domar, E. D. (1961). On the measurement of technological change. *The Economic Journal 71*(284), 709–729.

Edmond, C., V. Midrigan, and D. Y. Xu (2023). How costly are markups? *Journal of Political Economy 131*(7), 1619–1675.

Evans, M. D. D. and R. K. Lyons (2002). Order flow and exchange rate dynamics. *Journal of Political Economy 110*(1), 170–180.

Falvey, R. E. (1979). The composition of trade within import-restricted product categories. *Journal of Political Economy 87*(5), 1105–1114.

Feenstra, R. C. (1988). Quality change under trade restraints in Japanese autos. *The Quarterly Journal of Economics 103*(1), 131–146.

Feenstra, R. C. (1992). How costly is protectionism? *Journal of Economic Perspectives 6*(3), 159–178.

Foerster, A. T., P.-D. Sarte, and M. W. Watson (2011). Sectoral versus aggregate shocks: A structural factor analysis of industrial production. *Journal of Political Economy 119*(1), 1–38.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3),

733–772.

Glaeser, E. and J. Gyourko (2018). The economic implications of housing supply. *Journal of Economic Perspectives 32*(1), 3–30.

Grassi, B. (2017). IO in I-O: Size, industrial organization, and the input-output network make a firm structurally important. Working paper.

Gyourko, J. and J. Krimmel (2021). The impact of local residential land use restrictions on land values across and within single family housing markets. *Journal of Urban Economics 126*, 103374.

Harberger, A. C. (1954). Monopoly and resource allocation. *American Economic Review 44*(2), 77–87.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics 124*(4), 1403–1448.

Hsieh, C.-T. and E. Moretti (2019). Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics 11*(2), 1–39.

Hulten, C. R. (1978). Growth accounting with intermediate inputs. *The Review of Economic Studies 45*(3), 511–518.

Khandelwal, A. K., P. K. Schott, and S.-J. Wei (2013). Trade liberalization and embedded institutional reform: Evidence from chinese exporters. *American Economic Review 103*(6), 2169–2195.

Krueger, A. O. (1974). The political economy of the rent-seeking society. *American Economic Review 64*(3), 291–303.

Lipsey, R. G. and K. Lancaster (1956). The general theory of second best. *The Review of Economic Studies 24*(1), 11–32.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics 134*(4), 1883–1948.

McKenzie, L. W. (1951). Ideal output and the interdependence of firms. *The Economic Journal 61*(244), 795–803.

Peters, M. (2020). Heterogeneous markups, growth, and endogenous misallocation. *Econometrica 88*(5), 2037–2073.

Petrin, A. and J. Levinsohn (2012). Measuring aggregate productivity growth using plant-level data. *The RAND Journal of Economics 43*(4), 705–725.

Reischer, M. (2019). Finance-thy-neighbor: Trade credit origins of aggregate fluctuations. Working paper.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–720.

Rubbo, E. (2023). Networks, Phillips curves and monetary policy. *Econometrica 91*(4),

1417–1455.

Weitzman, M. L. (1974). Prices vs. quantities. *The Review of Economic Studies 41*(4), 477–491.