PRELIMINARY AND INCOMPLETE: DO NOT CITE

Misallocation in the Presence of Multiple Production Technologies*

Jose Asturias
Georgetown University Qatar

Jack Rossbach
Georgetown University Qatar

February 15, 2017

Abstract

While firms may employ different production technologies for producing the same or similar goods, the production technology used by each firm is typically not available in the data or otherwise known. Due to this limitation, researchers are often forced to assume all firms within an industry share a single production function. We develop a methodology based on cluster analysis that requires only firm level data on revenues and factor input expenditures, and allows us to identify when multiple production technologies are present in an industry and to identify the production technologies that each firm employs. We apply our methodology to Chilean plant level data and find strong evidence of multiple production technologies in most industries. We then evaluate the quantitative impact of this finding as it relates to the misallocation literature. While most studies of misallocation attribute all variation in factor input expenditure shares across firms within an industry to misallocation, we are able to use our framework to separate variation arising due to misallocation from variation resulting from differences in production technologies across firms. After accounting for the presence of multiple production technologies, we find that the estimated gains in manufacturing TFP and output from eliminating misallocation decrease by approximately 60 percent.

JEL classifications: D24, O11, O47

Keywords: Production Function Estimation, Cluster Analysis, Misallocation

*Asturias: jga35@georgetown.edu; Rossbach: Jack.Rossbach@georgetown.edu.

1 Introduction

It has long been understood that most goods can be produced using a variety of production technologies, each of which may require different combinations of factor inputs to produce the same output. One apparel factory might use a labor intensive technology where each garment is constructed by hand; whereas another apparel factory might make use of a capital intensive technology where garments are produced by a fully mechanized assembly line. Despite this wide-spread acknowledgement, that, for example, goods could be produced using a capital intensive technology or a labor intensive technology, this is rarely accounted for in macroeconomics papers and the macroeconomic implications are not well understood quantitatively. The reason relatively few papers have addressed this topic is not due to a lack of interest or presumed importance, but rather due to lack of data. While many data sources classify firms by their industry or by the products they produce, rarely are firms classified by the technology they use in production. Empirically, therefore, it is seen as difficult or impossible to identify the various technologies used within within an industry and to classify firms according to these technologies, except, perhaps, in studies focusing on a single industry where the various production technologies are already well understood ex-ante. For more general questions that focus on many industries simultaneously, however, there are few, if any, existing methodologies to identify heterogeneous production technologies used within an industry, or even whether there are, in fact, multiple production technologies in use by existing firms or not.

In this paper, we overcome this difficulty by developing a methodology to identify the number of production technologies within an industry, and to identify which technology is used by each firm, using observed variation in revenue cost shares for each factor input (total costs for that factor input divided by gross revenues) across firms. Our methodology is based on cluster analysis, which works by grouping observations so that they are similar within groups and dissimilar across groups, and is common in applications where pattern recognition is important, for example image processing and population inference based on genotype data. While the cluster analysis itself is non-parametric, we motivate its use through a structural model, which distinguishes our methodology from previous applications of cluster analysis in economics. In particular, our model allows us to show how the cluster analysis should be applied to the data, how to intpret the results, and under what conditions our methodology will successfully identify heterogeneous production technologies. The most important implication for our application of cluster analysis is that we are only able to identify production technologies that are sufficiently distinct in terms of their relative factor input intensities, relative to variation in observed revenue cost shares across firms within each production technology. This means that we are limited to only being able to identify a relatively small number of sufficiently distinct production technologies (or classes of production technologies), and are potentially understating the true number of production technologies. For example, while our methodology can be expected to identify whether some firms employ a labor intensive technology

and others employ a capital intensive technology, it would not necessarily be able to distinguish between two different labor intensive technologies. While future research should consider how to distinguish more similar production technologies, we do not consider this limitation a serious issue for this paper, as we want to be conservative when making the assertion that firms use different production technologies.

After verifying the validity of our methodology using simulated data, we apply our methdology to Chilean plant-level manufacturing data and we find strong evidence of multiple production technologies in the majority of industries. An important question is whether the different production technologies our clustering methodology identifies are in fact different production technologies. To justify our interpretation of the results, we show that our methodology works in a case where we know that firms employ different production technologies by showing that we are able to successfully identify firms from different industries (results not yet in paper). To further justify our interpretation, we conduct a case study on the copper industry (results not yet in paper). We show that our finding of multiple production technologies in the Chilean copper industry, and the characteristics of the firms that use each technology, align with industry reports describing the firms and production technologies used in the copper industry. While this evidence supports our interpretation of identifying different production technologies used by firms, our methodology could also be used to capture other sources of systematic differences in factor input revenue shares between groups of firms in an industry; for example, a policy that subsidizes capital rental rates for state owned enterprises relative to non-state owned enterprises.[1] While the paper proceeds assuming we have, in fact, identified differences in production technologies, in the appendix we show how our framework can be adapted to understand the quantitative impact of these alternative interpretations of our clustering results.

After establishing the presence of multiple production technologies in most industries, we evaluate the quantitative implications of this finding as it applies to the misallocation literature. Economists have made significant advances over the past decade in understanding how misallocation of inputs within and across firms can be a major factor in explaining the gaps in output and total factor productivity (TFP) between rich and poor countries. To identify misallocation, Hsieh and Klenow (2009) require micro-level data on the observed expenditures on inputs across and within individual firms, which is the same data we require to identify production technologies, as well as a way to measure how far away these observed allocations in the data are from the ideal or efficient allocation of inputs. Estimating the efficient allocation of inputs requires researchers to specify a production function for each firm in the economy, and the standard method for doing

[1] Our methodology is only able to distinguish systematic variation from idiosyncratic variation, and cannot further disentangle sources for systematic differences. If state owned enterprises both face a common subsidy to their capital rental rates and employ a different production technology than non-state owned enterprises, we would not be able to say what fraction of variation in factor input allocations is due to the subsidy and what fraction is due to differing production technologies.

so is to assume that production functions are identical for all firms within each industry, except for a productivity parameter that is allowed to differ across firms. Even within narrowly defined industries or product codes, firms are observed to have significant variation in their observed revenue cost shares; and misallocation is measured by assuming that the reason for this variation is unobserved distortions on the factor input prices that firms face, which leads to firms with different distortions to optimally choose different factor allocation bundles (conditional on the distortions facing all firms).[2] To the extent that these distortions could be removed, this leads to misallocation of factor inputs across firms, decreasing aggregate output and total factor productivity. Our methodology, however, allows us to loosen this restriction, and to estimate production functions separately for firms employing different production technologies. This is especially important for measuring misallocation, since if firms with different production technologies are assumed to have the same production function, we would incorrectly classify all differences in factor input shares as misallocation, instead of taking into account that some of the differences arise optimally due to differing production technologies. We find that accounting for variation due to the presence of multiple production technologies reduces the standard deviation of model implied distortions by an average of 16 percent across all inputs, and by an average of 34 percent for the input with the largest factor intensity in each industry.

We conduct a misallocation exercise similar to Hsieh and Klenow (2009) where we ask what the counterfactual gains in total output and TFP would be if we could completely eliminate the distortions on factor input prices that firms face and reallocation inputs across firms. We conduct this exercise both under the standard assumption that all firms in each industry share a single production function and then again with our methodology that can identify production technologies within industries. We find that accounting for the presence of multiple production technologies decreases the predicted gains in total output and TFP of eliminating misallcation by 60.8 percent. This indicates that accounting for multiple production technologies is important for understanding the potential gains from reducing misallocation. While we do not attempt to identify the sources of the unobserved distortions in this paper, we beleive our methodology may also be useful for researchers attempting to understand the origins of misallocation across firms. Previous research has largely failed to identify policies that can explain the model implied distortions observed in the data, however, our findings imply that, due to the misspecification of production functions, many of the firms previously thought to be undistorted, may in fact be highly distorted; while other firms thought to be highly distorted may in fact face relatively small distortions. It is our

[2]This method of assuming that all variation in factor revenue shares across firms is due to unobserved distortions means that other sources of variation, for example model misspecification and data mismeasurement, are absorbed into what we call misallocation. The question of how true misallocation can be distinguished from these other sources of variation is a topic of great interest. While our paper does not fully disentagle all sources of variation, it is the first paper we know of that attempts to answer what fraction of variation in the data can be explained by differences in production functions across firms.

hope, therefore, that by better allowing researchers to identify the distortion facing each firm, our methodology may be used to help find specific policies that can be targeted to reduce misallocation.

2 Model

In this section, we present the theoretical framework we will use to motivate our application of cluster analysis for identifying production technologies. In section 2.1 we present an economy with multiple production technologies without worrying about which features of the economy are observable in the data, and then in section 2.2 we specify which parameters of the economy are observable in the data and lay out our strategy for identifying the unobservable parameters.

2.1 Preliminaries

The framework we use is a standard model of monopolistic competition with heterogeneous firms. Since identifying production technologies within an industry will only require data for that industry, we focus this section on firms within a single industry (we consider a multi-industry extension when performing counterfactuals). The industry contains M firms, where each firm produces a differentiated product according to a constant returns to scale Cobb-Douglas technology. Where we depart from the standard framework of Hsieh and Klenow (2009) is that we allow for an arbitrary number of inputs, I, as opposed to only capital and labor, and we allow for multiple production technologies. Within the industry, there are H total production technologies, and the production function for firm m utilizing production technology h is

$$y_m = z_m \prod_{i=1}^{I} \left(x_m^i\right)^{\theta_h^i}, \tag{1}$$

where y_m is the firm's output, z_m its productivity, x_m^i is the quantity of input i, and θ_h^i is the factor intensity for input i by firms using technology h. Since production technologies exhibit constant returns to scale, therefore $\sum_{i=1}^{I} \theta_h^i = 1$ for all technologies h. The key innovation in our model is that θ_h^i will differ across technologies within an industry.

The industry has a perfectly competitive bundler which combines the differentiated outputs into an industry output according to a CES aggregator

$$Y = \left(\sum_{m=1}^{M} (y_m)^{\rho}\right)^{1/\rho}, \tag{2}$$

where $1/(1-\rho)$ is the elasticity of substitution across differentiated inputs. This implies that firm m faces demand for its output given by

$$c_m = \frac{E}{(p_m)^{\frac{1}{1-\rho}} (P)^{\frac{-\rho}{1-\rho}}}, \tag{3}$$

where E is total expenditures on the industry output (considered to be constant, and invariant to the distortions that will be defined below), p_m is the price of firm m's output, and P is the industry price index given by

$$P = \left(\sum_{m=1}^{M} (p_m)^{\frac{-\rho}{1-\rho}} \right)^{\frac{1-\rho}{-\rho}}. \tag{4}$$

Firms face distortions that lead to variation in the input prices they face. In particular, firm m faces an effective input cost of $\tau_h \tau_m^i p^i$ for input i, where p^i is the industry price of the factor input, τ_m^i is the idiosyncratic distortion firm m faces for input i, and τ_h is a distortion common for all firms employing production technology h, and which does not depend on the specific input. The presence of these distortions leads to dispersion in the marginal revenue products of inputs across firms within each in industry and leads to variation in the factor input choices of firms, even for firms that employ the same production technology. When we perform counterfactuals, these distortions will be the sole source of misallocation. The τ_h distortion can be thought of as a flat proportional tax or subsidy to the total costs firms face based on their production technologies, and, unlike the idiosyncratic distortions, is not tied to the firm's particular input usage. It is also important to note that in our framework, while p^i represents the price an undistorted firm will face for input i, this input price will still be determined in equilibrium and its value will depend on the distortions in the economy and will not, in general, be equal to the input price in a world with no distortions.

The firm solves the following profit maximization problem, taking as given the demand for their output and effective input prices,

$$\max \quad p_m y_m - \tau_h \sum_{i=1}^{I} \tau_m^i p^i x_m^i, \tag{5}$$

which has the solution that firms will charge a constant markup, $1/\rho$, over their marginal cost

$$p_m = \frac{1}{\rho} \frac{1}{z_m} \tau_h \prod_{i=1}^{I} \left(\frac{\tau_m^i p^i}{\theta_h^i} \right)^{\theta_i^{a(m)}}, \tag{6}$$

and implies that

$$\frac{p_m y_m}{\tau_h \sum_{j=1}^{I} \tau_m^j p^j x_m^j} = \frac{1}{\rho} \tag{7}$$

Note that the marginal cost for firm m depends on the firm's productivity, the firm's distortions, and the firm's production technology, and that the impact a distortion will have on a firm's marginal cost depends on the firm's production technology. Cost minimization implies that the share of each firm's total expenditures on any given input and including disotrtions, will be equal to the firms factor intensity for that input

$$\frac{\tau_m^i p^i x_m^i}{\sum_{j=1}^I \tau_m^j p^j x_m^j} = \theta_h^i. \tag{8}$$

The importance of this equation is that it highlights that all firms with the same production technology will have the same the same expenditure shares on each input when distortions are included. In particular, it relates distortions and input expenditures, to factor intensities, a relationship we will ultimately use to identify the production technologies in each industry. It also makes it clear that we should expect some variation in input cost shares to arise even without distortions as long as firms employ different production technologies, indicating that this variation would be misclassified by a framework that only allowed for a single production function.

2.2 Connecting the Model to Data

Up until this point, we haven't worried about which features of the model are typically available in the data or not. The reason for this paper, however, is that many of the parameters in section 2.1 are not available in standard data, and that this unavailbility is the reason most papers are forced to assume a single production technology is used in each industry. For example, if factor intensities were directly observable, then we would instantly know whether different firms were employing different production technologies. Likewise, if we were able to observe each firm's input expenditures with the distortions included, then we would be able to use equation 8 to recover the factor intensities and therefore the production techhniques.

Instead, we operate under the assumption that the deep parameters of the model, in particular markups and factor intensities, are not observable; and that the input expenditures we observe exclude the idiosyncratic and technology specific distortions.[3] While some data breaks down values into prices and quantities, many datasets contain only revenues and input expenditures without providing quantity measures for either. Therefore we assume that the only features of the model we are able to observe in the data are distortion excluded expenditures on each input, i.e. $p^i x_m^i$, and total revenues, $p_m y_m$, for each firm in the industry. It is important to understand that even though we don't observe the distortions, the distortion excluded expenditures and firm revenues are equilibrium objects which will depend on the unobserved distortions and are not equal to expenditures and revenues in a world without distortions. We define the (distortion excluded) revenue cost share for input i, $\tilde{\theta}_{i,m}^s$, to be observed (distortion-excluded) input expenditures divided by revenues, i.e $p^i x_m^i / p_m y_m$. We can then combine equation 8, which relates input expenditures with distortions and factor intensities, with equation 7, which relates input expenditures to the

[3]How do we know the input expenditures exclude distortions? Precisely because we observe significant variation in revenue cost shares across firms. The only way to justify this if we observe distortions in our input expenditures is if each firm has it's own production technology. In this case there would be no way to no way to identify idiosyncratic distortions from idiosyncratic differences in . whether a firm's allocation is optimal or not and perform counterfactuals for misallocation.

markup over marginal cost, to derive the following expression for the revenue cost share for firm m for input i

$$\tilde{\theta}_m^i = \frac{\rho\theta_h^i}{\tau_h\tau_m^i}. \tag{9}$$

This equation relates each firm's distortion-excluded expenditures on an input to its technology specific distortion, its idiosyncratic distortion for the markup, and factor intensity. This expression will the the main equation we use throughout the paper to identify unobservable parameters and, eventually, identify production technologies. The key benefit of using revenue cost shares is that the revenue cost share for a given input does not depend on the factor intensities or distortions for other outputs. On the other hand, equation 9 makes it clear that if we were to simply use observed input cost shares, $p^i x_m^i / \sum_{j=1}^I p^j x_m^j$, then the cost share for each input would depend on all of the idiosyncratic distortions simultaneously.

Our ultimate strategy for recovering production technologies and other unobserved parameters is to use variation in revenue cost shares for each input to pin down factor intensities and distortions. The key insight that will allow us to decompose the right hand side of equation 9 is that, while there are four unknown objects that can capture variation in revenue cost shares, the unknowns all vary differently across firms, inputs, and production technologies. In particular, idiosyncratic distortions for each input vary by firm; factor intensities for each input vary by production technology; the technology specific distortion varies by production technology, but does not vary across inputs; and the markup is common for all firms in an industry. Therefore, if we know which firms share a common production technology, then it is relatively straightforward to estimate the factor intensities for each technology, as well as markups and distortions. Since we can't directly observe which firms share a common production technology, in the next section we lay out our methodology for using variation in revenue cost shares to identify when multiple production technologies are employed within an industry and to group firms that share a common production technology.

3 Recovering Production Technologies

In this section we lay out our methodology for identifying the production technologies used by firms within an industry. Our methodology will be based off of cluster analysis, which is a set of technologies used to group observations so that they are similar within groups and dissimilar between groups. Our application of cluster analysis will be based off the expression for revenue cost shares derived in the previous section and will be based on the intuition that, in an industry with multiple production technologies, we should expect gaps in the dispersion of revenue cost shares to arise between sets of firms with different production technologies. Even when the variance of distortions is high, we would still expect to encounter areas of higher and lower density in the

distribution of revnue cost shares, where high density areas indicate firms that share a common production technology and low density areas indicate differences in factor intensites for firms with different production technologies.

We present our methodology in several steps. In a preliminary step, we motivate the variables in the data that we will use in our cluster analysis. In steps 1–3 we use a clustering methodology to find the number of technologies for each industry, $\hat{H}$, and a partition that assigns each firm to a technology, $\hat{a}_{\hat{H}}(m)$, given $\hat{H}$ technologies. In steps 4–6, we show how to exploit the intuition in section 2.2explain the identification of ρ, the factor intensities of each technology, and the firm-level distortions. In section 3.9, we use Monte Carlo simulations to show that our methodology performs well in recovering both the number of technologies and the assignment of firms to these technologies.

3.1 Preliminaries

The firm's m FOC yield to the following expression for firm m:for each input i. The left-hand side of equation 9 is the observed share of expenditures on input i as a fraction of revenue, which we denote as $\tilde{\theta}^i_m$.[4] This expression is useful for two reasons. First, we can usually calculate $\tilde{\theta}^s_{i,m}$ using micro-level datasets since it only involves the total cost for each input and revenue. Second, for all firms using the same technology, all variation in $\tilde{\theta}^s_{i,m}$ arises due to idiosyncratic distortions within the lens of the model.

We take the log of both sides of equation 9 and we get

$$\log \tilde{\theta}^s_{i,m} = \log \rho^s + \log \theta_i^{a(m)} + \log \tau^s_{y,a(m)} - \log \tau^s_{i,m}. \tag{10}$$

Equation 10 shows that, within an industry, there is no way for us to distinguish between differences in $\tau^s_{i,m}$ and differences in θ^h_i. This is the usual fundamental identification problem of separating differences in technology from differences in distortons. In the literature, it is usally assumed that technology does not vary within an industry. Here, we relax this assumption but still need to make an identifying assumption about how distortions are distributed. In particular, we assume that $\log \tau^s_{i,m}$ is normally distributed such that the distribution of $\left(\tau^s_{i,m}\right)^{-1}$ for a given technology h and input i has mean 1. The distribution of $\log \tau^s_{i,m}$ is also characterized by a variance-covariance matrix Ω^s_h for each technology h. Note that this allows for correlated distortions across inputs of firms with the same technology. In the application of the methodology, we show that the assumption of $\log \tau^s_{i,m}$ being normally distributed is consistent with the data.

[4]Note that equation 9 is similar to equation 18 in Hsieh and Klenow (2009), which was used to identify their output distortion.

3.2 Step 1: Clustering Firms by Production Technology

The next step is to divide firms in a sector according to their production technology. In this step, we apply the k-means clustering methodology assuming that there are H' technologies. We repeat this process for $H' = 1, ..., H'_{\max}$ technologies, where $H'_{\max}$ is the maximum number of production technologies considered. The process yields partition $\hat{a}_{H'}(m)$ that assigns firms to technologies for all H'. In steps 3 and 4, we describe the selection of the number of clusters.

Given our assumption that there are H' technologies in a sector, we want to group the logged expenditure shares characterized in equation 10 in a way that we maximize the similarity within each group. To do so, we partition the set of firms so that we minimize the sum of squared errors (SSE) between the logged expenditure shares of firms and the mean across firms using the same technology. We define $SSE_{H'}$ to be

$$SSE^s_{H'} = \min_{\{\hat{a}^s_{H'}(m)\}} \sum_{h'=1}^{H'} \sum_{m=1}^{M^s} \sum_{i=1}^{I} \mathbb{I}_{h'=\hat{a}^s_{H'}(m)} \left(\log \tilde{\theta}^s_{i,m} - \xi^s_{i,h'}\right)^2, \tag{11}$$

where $\mathbb{I}_{h'=\hat{a}^s_{H'}(m)}$ is an indicator function which takes on the value of one if firm m is assigned to technology h', $\xi^s_{i,h'}$ is the mean of the logged expenditure share of input i in technology h'. Notice that $SSE^s_{H'}$ uses the $\hat{a}^s_{H'}(m)$ partition that minimizes the SSE.

Finding the globally optimal partition to solve the problem in equation 11 is challenging due to the high dimensionality of the problem. Even with just two technologies and 50 firms, there are 2^{50} possible combinations. To tackle this problem, we use k-means, which is a commonly used algorithm for this type of problem. To implement this methodology, we first select H' "centers." A center is a set of input intensities of a given technology. The standard algorithm randomly selects observations with equal probability as the starting point. We then follow a two-step procedure: 1) classify: we assign each firm to the closest center, 2) recenter: once we have assigned firms to the closest center, we set the new center to be the component-wise mean of all the observations assigned to that technology. We then repeat steps 1 and 2 until the algorithm converges.[5]

It is important to note that if the initial starting point is not close enough to the solution, k-means may converge to a local minima. Thus, it is standard to re-initialize the algorithm many times with different starting points, and then selecting the partition with the lowest SSE. We also use the k-means++ methodology developed by Arthur and Vassilvitskii (2007) to select starting points in a way that improves the speed and accuracy of the algorithm. In this method, the first center is randomly selected among the observations. For the following centers, the probability of selecting an observation is proportional to its distance from the previously chosen center.

[5] Another way of understanding k-means is that the goal of the algorithm is to find assignments of firms and centers in such a way as to minimize the SSE. In the classify stage, we take the centers as given and assign firms to the closest center, which minimizes SSE. In the recenter stage, we take firm assignments as given and set the new centers to equal the mean of the firms in each cluster, which also minimizes SSE.

3.3 Step 2: Determining Whether There is More Than One Technology

We now determine whether there is evidence that is consistent with more than one technology being present in a sector. We use our measure of model fit, $SSE_{H'}$, to make this determination. It is important to note that $SSE_{H'}$ always decreases as the number of technologies increases. Consider the extreme case in which H' is equal to the number of firms in the sector. Then, we would have a perfect fit of the data and $SSE_{H'} = 0$. Therefore, it is important to appropriately penalize declines in $SSE^s_{H'}$ associated with larger H'.

To penalize the $SSE^s_{H'}$ for the case in which we have H' technologies relative to one technology, we numerically calculate the maximal reduction in SSE. The maximal reduction is the percent reduction in SSE of going from one to H' technologies even though the underlying data was generated with only one technology.

To determine the maximal reduction, we begin with the assumption that a sector has only one technology. We fit a lognormal multivariate distribution to the logged expenditure shares. To do so, we compute the average of the logged expenditure share for each input to find the mean. We also compute the variance-covariance matrix of the logged expenditure shares across inputs.

Using this distribution, we generate 100 separate synthetic datasets in which the number of draws equals the number of plants in the industry. We run the k-means clustering algorithm on synthetic dataset j and calculate $SSE^{\text{Synthetic}}_{H',j}$ for $H' = 1, ..., H'_{\max}$. We define the maximal reduction in synthetic dataset j when we utilize H' technologies to be

$$R^s_{H',j} = \left| \frac{\widetilde{SSE}^s_{H',j} - \widetilde{SSE}^s_{1,j}}{\widetilde{SSE}^s_{1,j}} \right|, \tag{12}$$

where $\widetilde{SSE}^s_{H',j}$ is the SSE of synthetic dataset j with H' technologies. Let $\hat{j}^s(H')$ be the synthetic dataset in which the maximal reduction is in the 95th percentile among cases in which we use H' technologies. We calculate the maximal reduction under multiple synthetic datasets in order to capture the fact that this statistic can change depending on the particular draws of input distortions. This is especially important since most sectors in our data application have less than 150 plants. Note that by using the 95th percentile, we are choosing a relatively high threshold that must be met in the data.

We conclude that there are more than one production technologies if

$$\left| \frac{SSE^s_{H'} - SSE^s_1}{SSE^s_1} \right| > R^s_{H',\hat{j}^s(H')}, \tag{13}$$

for any $H' > 1$. Thus, if the condition in equation 13 is not satisfied, we assume that there is only one technology in the industry.

3.4 Step 3: Selecting the Number of Technologies

After concluding that there are multiple production technologies in a sector, we consider two different methods to determine the number of technologies. This can be thought of as first testing the null hypothesis that there is only a single production technology. Then, conditional on rejecting the null hypothesis, we choose the number of technologies $H \geq 2$. We follow this two-steps process since the methodologies that we now describe assume that there are at least two technologies.

The first selection criteria we consider is the CH index by Calinski and Harabasz (1974), which is the most commonly used criterion for determining the number of clusters. The CH index for H' technologies is defined as

$$CH^s_{H'} = \frac{SST^s - SSE^s_{H'}}{SSE_{H'}} \frac{N^s - H'}{H' - 1}, \tag{14}$$

where SST is the total sum of squared errors. The expression for SST^s is

$$SST^s = \sum_{m=1}^{M^s} \sum_{i=1}^{I} \left(\log \tilde{\theta}^s_{i,m} - \xi^s_i \right)^2, \tag{15}$$

where ξ_i is the mean of the logged expenditure share for input i of all firms, regardless of technology. Note that SST^s characterizes the total variation in the data and thus does not depend on H'. After computing the CH index for each $H' \geq 2$, we choose the number of technologies that maximizes the index.

To understand how the CH index works, note that the first term of equation 14, $(SST^s - SSE^s_{H'})/SSE^s_{H'}$, measures the overall fit of the model and increases as $SSE_{H'}$ declines. Thus, the first term increases as H' increases. The second term, $(N^s - H')/(H' - 1)$, is a penalty term, which decreases as more technologies are added. Thus, the CH index increases only when the improvement in fit from a larger H' overcomes the penalty term.

The other selection criteria that we consider is the Lower Bound technology (LBT) by Steinley and Brusco (2011), a more recently developed selection criterion that more fully utilizes the structure of the data. The LBT index is computed as

$$LBT_{H'} = \frac{SSE^s_{H'} - SSE^s_{LBT,H'}}{SST^s}. \tag{16}$$

The expression for $SSE^{LBT}_{H'}$ is

$$SSE^s_{LBT,H'} = \text{trace}(X'X) - \sum_{j=1}^{H'} \lambda^s_j, \tag{17}$$

where X is a matrix with M rows and I columns where each row is an firm in the data and the columns are the logged expenditure shares, and λ_j is the jth largest eigenvalue of XX'. $SSE^s_{LBT,H'}$ is a non-binding lower bound for $SSE_{H'}$. After computing the LBT index for each $H' \geq 2$, we choose the number of technologies that minimizes the index. For LBT to work properly, we must

have more inputs (columns) than technologies. For this reason, in our application to Chilean plant-level data in section 4, we will consider four inputs (unskilled labor, skilled labor, capital, and intermediate inputs) and a maximum of three technologies.

The LBT method has the advantage of taking the structure of the data into account when creating a penalty term. In contrast, the CH index only considers the number of observations. The CH and LBT model selection criteria usually agree on the same number of production technologies in the Monte Carlo simulations that we present in section 3.9. In general, the CH model selection criteria performs better when the distribution of distortions is similar across all inputs and the LBT criteria performs better when the distribution of distortions is asymmetric across inputs. For example, the LBT method tends to perform better if there is considerably more variation across distortions for capital compared to unskilled labor. To ensure that our results are as conservative as possible, we set $\hat{H}$ to be the minimum number of production technologies predicted by the two criteria when they do not coincide. We find that this works better than using either of the penalty criteria alone.

3.5 Step 4: Identifying ρ^s

For every sector, we have now determined the number of technologies, $\hat{H}(s)$, and the assignment of each firm to a technology, $\hat{a}_{\hat{H}(s)}(m)$. We now use the structure of the model to determine ρ^s for each sector. First, we use equation 9 to find the following relationship for input i of firms using technology h

$$\frac{\sum_{m=1}^{M} \mathbb{I}_{h=\hat{a}^s_{\hat{H}(s)}(m)} \tilde{\theta}^s_{i,m}}{N^s_h} = \hat{\rho} \theta^h_i \hat{\tau}^s_{y,h} \frac{\sum_{m=1}^{M} \mathbb{I}_{h=\hat{a}^s_{\hat{H}(s)}(m)} \left(\hat{\tau}^s_{i,m}\right)^{-1}}{N^s_h}. \tag{18}$$

where $N^s_h = \sum_{m=1}^{M^s} \mathbb{I}_{h=\hat{a}^s_{\hat{H}(s)}(m)}$ is the number of firms that use technology h under $\hat{a}^s_{\hat{H}(s)}(m)$. We use the condition that for any input i in technology h

$$\frac{\sum_{m=1}^{M} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \left(\hat{\tau}^s_{i,m}\right)^{-1}}{N^s_h} = 1, \tag{19}$$

which is the mean of the distribution of the distortions described in section 3.1. We use the condition in equation 18 and add up across all inputs for firms that use technology h to find

$$\sum_{i}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \tilde{\theta}^s_{i,m}}{N^s_h} = \hat{\rho}^s \hat{\tau}^s_{y,h} \sum_{i}^{I} \hat{\theta}^h_i. \tag{20}$$

We use the constant returns to scale condition, $\sum_i \theta^h_i = 1$. We use the condition in equation 20 and sum over all technologies

$$\hat{\rho}^s \sum_{h}^{\hat{H}(s)} \hat{\tau}^s_{y,h} = \sum_{h}^{\hat{H}(s)} \sum_{i}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \tilde{\theta}^s_{i,m}}{N^s_h}. \tag{21}$$

We use the condition that $\sum_h \hat{\tau}_{y,h}/\hat{H}(s) = 1$ to arrive at

$$\hat{\rho}^s = \frac{1}{\hat{H}(s)} \sum_{h=1}^{\hat{H}(s)} \sum_i^I \frac{\sum_{m=1}^{M^s} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \tilde{\theta}^s_{i,m}}{N^s_h}, \tag{22}$$

which allows us to determine ρ.

To understand the identification strategy, consider an industry in which we observe that expenditures on inputs is a small fraction of revenue compared to other industries. Equation 22 implies that ρ will be low, which is consistent with high markups in the industry.

3.6 Step 5: Identifying Technology-Specific Distortions

We can now determine the technology-specific distortion. To do so, we re-arrange equation 20 as follows

$$\hat{\tau}^s_{y,h} = \frac{1}{\hat{\rho}^s} \sum_i^I \frac{\sum_{m=1}^{M^s} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \tilde{\theta}^s_{i,m}}{N^s_h}, \tag{23}$$

which allows us to determine this parameter using $\hat{\rho}^s$ from step 4.

3.7 Step 6: Identifying Factor Intensities for Each Technology

The next step is to identify the factor intensity of the production function, θ^h_i, for input i and technology h. To do so, we manipulate the condition in equation 18 as follows

$$\hat{\theta}^h_i = \frac{1}{\hat{\tau}^s_{y,h}\hat{\rho}^s} \frac{\sum_{m=1}^{M^s} \mathbb{I}_{h=\hat{a}^s_{\hat{H}}(m)} \tilde{\theta}^s_{i,m}}{N^s_h}, \tag{24}$$

which allows us to determine the factor intensities given from $\hat{\tau}^s_{y,h}$ and $\hat{\rho}^s$ before. The intuition is that we can take the average expenditure on an input as a fraction of revenue for all firms using a technology. The only adjustment, reflected in ρ^s, is to adjust for markups which are included in the value of revenue.

3.8 Step 7: Identifying Firm-Level Distortions

The last step is to determine the firm-level distortions for firm m and input, $\hat{\tau}^s_{i,m}$. To do so, we re-arrange equation 9 as follows

$$\hat{\tau}^s_{i,m} = \frac{\hat{\rho}\hat{\theta}^{\hat{a}_{\hat{H}}(m)}_i \hat{\tau}_{y,a(m)}}{\tilde{\theta}^R_{i,m}} \tag{25}$$

This equation allows us to find the distortion for each firm using $\hat{\rho}^s$, $\hat{\tau}^s_{y,h}$, and $\hat{\theta}^h_i$ from the previous steps.

3.9 Monte Carlo Exercise

The goal of this section is to examine how our methodology performs in recovering heterogenous production technologies within a sector. To do so, we simulate data from the model and then use the clustering methodology to recover production technologies and distortion parameters in the simulated data. We find that the clustering methodology performs well in recovering firm-level technologies.

3.9.1 Simulation Strategy

Each sector has either one, two, or three technologies. We select the number of technologies to be uniformly distributed across these choices. Each production technology has four inputs and the factor intensities vary across these technologies.

To create a production technology, in the first step we decide whether a factor is used intensively. We assign an input as being used intensively with a 50 percent probability. In the second step, we select from a uniform distribution of [1/3, 1] if the factor is used intensively and from a unform distribution of [0, 1/3] if the factor is not used intensively. In the third step, we arrive at the factor intensities by re-scaling the selected numbers for all of the inputs so that they sum to one. If we have a case in which two technologies have four inputs with the same intensities in step one, then we drop that technology. This way of generating the production technologies ensures that the differences are large enough to be detected by the clustering methodology.

For each production technology, we use a uniform distribution to generate between 20 and 50 firms. For each firm, we generate a distortion, $\tau_{i,m}$, for each input i drawn from a multivariate lognormal distribution. The mean of the lognormal distribution of distortions for inputs is uniformly distributed between -0.2 and 0.2. The standard deviation is uniformly distributed between 0.1 and 0.3. We generate a correlation matrix for the input distortions such that the diagonal entries are 1 and the off-diagonal entries are uniformly distributed between -0.2 and 0.2.[6] The technology-specific distortion is uniformly distributed XXX. The last parameter generated is the elasticity of substitution, which is drawn so that markups are uniformly distributed across simulations between 15 and 100 percent. Note that the firm-level productivity, z_m, does not matter for any of the components of equation 10.

Table I summarizes the parameters used for the Monte Carlo exercise.

3.9.2 Simulation Results

After generating the synthetic data, we run all of the steps of the clustering methodology described in section 3. The methodology performs well in finding the true number of technologies and in

[6] Given the standard deviation of each input distortion and the correlation matrix across the inputs, the variance-covariance matrix can be determined.

Table I

Selection of Parameters for Monte Carlo Exercise

Parameter	Generating Process
Number of production technologies (H)	Uniform: 1, 2, and 3
Number of inputs (I)	4
Factor intensity for input i for production technology h (θ_i^h)	Uniform distribution over [0,1/3] and [1/3,1] with equal probability (re-scaled to sum to 1)
Number of firms assigned to production technology h	Uniform: [20,50]
Mean of lognormal distribution of input i distortion	Uniform: [-0.2,0.2]
Standard deviation of lognormal distribution of input i distortion	Uniform: [0.1,0.3]
Correlation matrix of distortions	Diagonal of one, other entries uniform: [-0.2,0.2]
Distortion of technology h	XXX
Parameter that governs elasticity of substitution (ρ)	Markup is uniform: [1.15,2]

Notes: Table I summarizes the data-generating process for the parameters of the Monte Carlo exercise. Column 1 shows the parameter of interest. Column 2 describes the generating process for that parameter.

assigning firms to the correct technology. Table II compares the true number of technologies (in the row) with the number of technologies we recover (in the column). We successfully recover the true number of technologies at a very high rate regardless of whether there are 1, 2, or 3 technologies (98.6-100 percent). Table III shows the percentage of firms that we correctly classify by technology. This table only considers the cases in which we correctly determine the true number of technologies. Firms are assigned to their true technology in 99.9-100 percent of the simulations.

Lastly, we evaluate how our methodology performs at recovering the other underlying parameters. We again focus on the cases where we correctly recover the true number of technologies. We construct the Average Absolute Percentage Error as follows:

$$\text{Absolute Percentage Error} \equiv \left| \frac{\Omega^{\text{true}} - \Omega^{\text{recovered}}}{\Omega^{\text{true}}} \right|. \tag{26}$$

The results for the factor intensities, distortions, and markups are reported below in Table IV. In all cases, the Average Absolute Percentage Error is less than 2 percent, indicating a high degree of accuracy in parameter recovery. For example, if the true factor intensity is 0.35, our results suggest that the recovered intensity will likely be between 0.34 and 0.36, centered around 0.35. If we allow positive errors to offset negative errors, and only compute the average percentage error without taking absolute values, then the errors are effectively zero, meaning our recovered parameters are centered around the true parameters.

Table II

True vs. Recovered Number of Production Technologies

	Number of Technologies Recovered			Percent
True Number of Technologies	1	2	3	Correct
1	349	0	0	100
2	2	360	3	98.6
3	0	4	282	98.6

Notes: Table II shows the true number of technologies vs. the number of production technologies recovered in the Monte Carlo exercise. Column 1 describes the true number of production technologies in the simulated data. Columns 2-4 describe the number of simulations that recovered 1, 2, and 3 technologies respectively. Column 5 shows the percent of simulations in which we recovered the correct number of technologies.

Table III

Percentage of Firms Correctly Classified by Technology

	True Number of Technologies		
	1	2	3
Percent Correct	100	99.9	99.9

Notes: Table III shows the percent of firms that were correctly classified by their technology use in the Monte Carlo exercise. The columns indicate the number of production technologies in the simulated data. In this table, we only report results for simulations in which we recover the correct number of technologies.

4 Application to Chilean Plant-Level Data

We now apply the clustering methodology developed in section 3 to a panel of Chilean plant-level data.

4.1 Data Description

We use the Encuesta Nacional de Industria Anual (ENIA) dataset collected by the Instituto Nacional de Estadisticas (INE). This dataset covers all manufacturing plants with more than 10 employees between the years of 1995-2006. The dataset uses ISIC revision 3 classification system at the 4-digit level. The ENIA Chilean plant-level data has been widely used in previous studies.[7]

[7]The ENIA plant-level data has been previously used by others. A few papers that have used Chilean plant-level data include: Alv (2005), Bergoeing and Repetto (2006), Pavcnik (2002), Petrin and Levinsohn (2012), Levinsohn (1999), Levinsohn and Petrin (2003).

Table IV

Average Absolute Percent Error between True and Recovered Parameters

	True Number of Technologies		
	1	2	3
Factor Intensity	1.7	1.6	1.7
Distortions	1.9	1.7	1.7
Markups	0.6	0.5	0.6

Notes: Table IV shows the Absolute Percentage Error (as described in equation 26) for factor intensity, distortions, and markups in the Monte Carlo exercise. Column 1 reports the variable of interest. Columns 2-4 indicate the true number of production technologies in the simulated data. In this table, we only report results for simulations in which we recover the correct number of technologies.

For output, we use the nominal gross output reported by plants. As inputs, we use capital, unskilled labor, skilled labor, and intermediate inputs. For capital, we use the nominal book value of capital. For skilled and unskilled labor, we use employee man-hours. For intermediate inputs we use nominal value reported by plants.

Next, we construct the nominal cost shares for each input at the plant-level. In the case of the intermediate inputs, this value is reported by the plants. For labor, we use labor remuneration for each type of labor. We do not have a direct measure of the user cost of capital. We utilize a strategy similar to Young (1995) to recover the use cost of capital by using the following no-arbitrage condition for each type of capital

$$R_{tj} = 1 + r_t - (1 - \delta_j) \frac{P_t}{P_{t+1}} \frac{P_{t+1}^K}{P_t^K}, \tag{27}$$

where R_{tj} is the user cost of capital for capital type j, P_t is the price level of the aggregate economy, P_t^K is the price of a unit of capital, and r_t is the real interest rate. We use the economy-wide real interest rate for r_t, the GDP deflator to determine P_t, and the investment deflator to determine P_{t+1}^K / P_t^K. We multiply the the user cost of capital for each type of capital, found using equation 27, by the nominal value of that capital.

Finally, we consider industries that have at least 50 plants. After removing outliers and industries with an insufficient number of firms, we are left with 13 industries containing 1360 firms.

4.2 Clustering Results

We apply the clustering methodology on the Chilean plant-level data by applying steps 1-7 described in section 3. The process yields the number of technologies in each industry, assignment of plants to the various technologies, ρ parameter for each industry, factor intensities for each

Table V

Number of Technologies Across Industries

ISIC	Industry Description	Count	Prod. Technologies
1511	Production, processing and preserving of meat and meat products	75	1
1512	Processing and preserving of fish and fish products	108	2
1513	Processing and preserving of fruit and vegetables	61	2
1531	Manufacture of grain mill products	78	2
1541	Manufacture of bakery products	383	2
1552	Manufacture of wines	67	2
1810	Manufacture of wearing apparel, except fur apparel	133	2
2010	Sawmilling and planing of wood	124	2
2102	Manufacture of corrugated paper and paperboard and of containers of paper and paperboard	50	2
2520	Manufacture of plastics products	194	2
2695	Manufacture of articles of concrete, cement and plaster	71	2
2811	Manufacture of structural metal products	84	2
2899	Manufacture of other fabricated metal products n.e.c.	92	2
3610	Manufacture of furniture	108	2

Notes: Table V shows the number of technologies determined by the clustering methodology at the industry-level. Column 1 is the ISIC code for the industry. Column 2 is a description of that industry. Column 3 lists the number of plants. Column 4 shows the number of production technologies.

technology, and distortions for each plant. We would like to better understand the implications of considering multiple production technologies. Thus, we repeat steps 4-7 under the assumption that there is only one production technology.

Table V reports the number of production technologies identified in each industry and the number of plants in the industry. Our results indicate that almost all industries have multiple technologies: 13 industries have 2 technologies present and 1 industry has only 1 technology. We find that incorporating these multiple technologies has an important impact on the distribution of the distortions. Table VI lists summary statistics for the change in the distortions of inputs when we consider more than one technology. Columns 2-5 list the percent decline in the standard deviation of the distortions for each of the four inputs. We find that there are consistent reductions in the standard deviation of the input distortions. We also find that the largest declines in standard deviation are in the distortions for capital, with an average decline of 21.8 percent across all industries. We find that the smallest declines are in intermediate inputs, with an average decline

of 5.4 percent.

We now focus on the results from the manufacturing of non-ferrous and precious metals industry (ISIC 2720) to get a better understanding of the clustering results. We focus on this industry since it is the largest by gross output. It is also one of the industries with the largest declines in the standard deviation of the distortions of inputs, as shown in Table V, thus clearly illustrating the results from the methodology. Table VII describes the recovered factor intensities under two technologies and under one technology. First, when we allow for two technologies, we find 23 plants are assigned to technology 1 and 34 plants are assigned to technology 2, reflecting widespread use of both technologies. Second, we find that when we use one technology, we tend to infer factor intensities in between those that we find when we have two technologies. For example, when we use two technologies we find that technology 1 has a factor intensity of 0.94 for intermediate inputs and technology 2 has a factor intensity of 0.77. On the other hand, if we only consider one technology then the inferred factor intensity is 0.82.

In Figure I, we plot the kernel density of the distortions for each input when we allow for multiple technologies or when we only allow for one technology. We tend to see large declines in the standard deviation of the distributions. These declines are particularly prominent when we examine the right-tail of the distribution for capital, skilled labor, and unskilled labor.

4.3 Gains from Eliminating all Distortions

In order to estimate the gains in manufacturing output and TFP from eliminating all distortions in the economy, we impose additional structure to the model presented in section 2. We consider a multi industry extension where the output from each industry (let the subscript s denote the industry for each variable from section 2) will be used to create a single final consumption good, Y, by a perfectly competitive bundler with a Cobb-Douglas production function.

$$Y = \prod_{s=1}^{S} (Y_s)^{\alpha^s} . \tag{28}$$

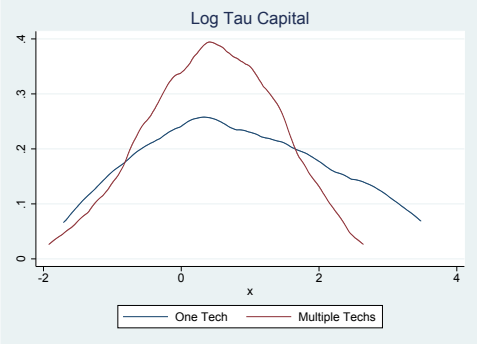
Here Y^s is the output from industry s, and α^s is equal to industry s's share of total expenditures in equilibrium, regardless of distortions. This multi industry framework allows us to consider how removing distortions would lead to realloction of inputs across industries as well as within industries. We assume that factors are fully mobile across industries, so that the prices of each input will be the same for all firms in the economy after removing distortions. The buget constraint of the representative consumer implies that total expenditures will equal income from factor inputs, profits from firms, Π, and transfers from firms due to distortions, T. Note that the transfers do not necessarily need to be government transfers due to explicit taxes or subsidies; consumers own the factors of production so the transfers are implicit based on the wedge between the undistorted

Figure I

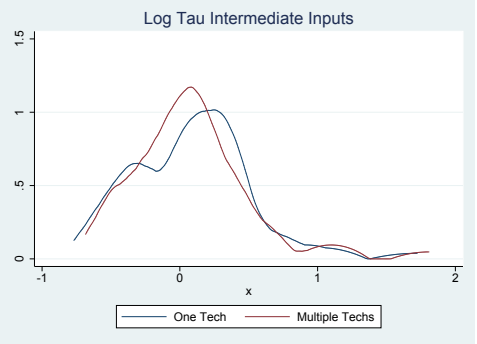
Kernel Density Plots of Logged Input Distortions

Manufacturing of Non-Ferrous and Precious Metals (ISIC 2720)

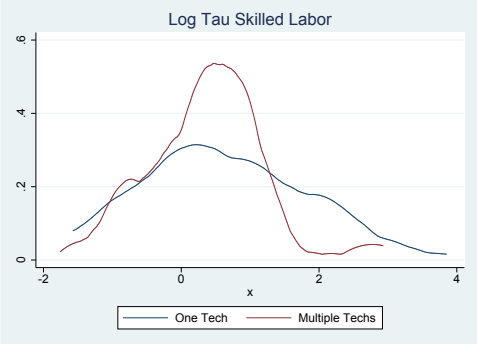
A: Capital



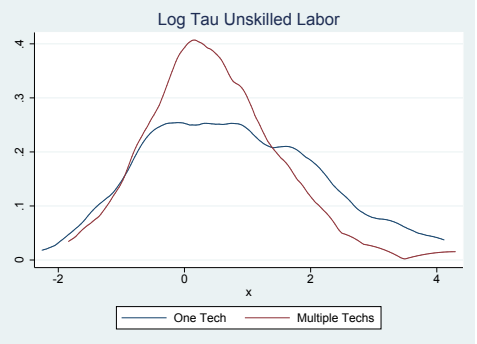
B: Intermediate Inputs



C: Skilled Labor



D: Unskilled Labor



Panel A of Figure I shows the kernel density plot of the logged input distortions of capital under one technology and multiple technologies; Panel B shows the same plot for the input distortions of intermediate inputs; Panel C shows the same plot for the input distortions of skilled labor; Panel D shows the same plot for the input distortions of unskilled labor.

price and what the consumer receives from the firm. Therefore total expenditures are equal to

$$PY = \sum_{i=1}^{I} p^i X^i + \Pi + T \tag{29}$$

where X^i denotes the exogeneously determined and inelastically supplied total amount of factor i in the economy. Profits are the result of markups over marginal cost (markups are allowed to vary across industries) and are given by

$$\Pi = \sum_{s=1}^{S} \frac{1}{\rho_s} \left(\sum_{m=1}^{M_s} p_{m,s} y_{m,s} \right), \tag{30}$$

while transfers are defined as

$$T = \sum_{s=1}^{S} \sum_{h=1}^{H_s} \sum_{m=1}^{M_{h,s}} \sum_{i=1}^{I} \left(\tau_{h,s} \tau_{m,s}^{i} - 1 \right) p^i x_{m,s}^{i}. \tag{31}$$

Note that since we can back out distortions and markups from the data using our methodology in section 3, we are able to use the above formulas to determine the initial equilibrium values of profits and transfers.

The clearing condition for each input i is

$$x_i = \sum_{s=1}^{S} \sum_{m=1}^{M^s} x_{i,m}^s. \tag{32}$$

We must now calibrate the productivities of each plant. To do so, we first normalize the productivity of one plant in industry s, $\tilde{m}(s)$, to 1. We then find the productivities for all other plants in the industry by matching the relative size of the gross output with the plant that we normalized the productivity for. We combine equations 3 and 6 to find the productivity of plant m relative to that of $\tilde{m}(s)$

$$\frac{z_m^s}{z_{\tilde{m}(s)}^s} = \left(\frac{p_m^s c_m^s}{p_{\tilde{m}(s)}^s c_{\tilde{m}(s)}^s} \right)^{\frac{1-\rho^s}{\rho^s}} \frac{\prod_{i=1}^4 \left(\frac{\tau_{i,m}^s p_i^x}{\theta_i^{a_s(m)}} \right)^{\theta_i^{a_s(m)}} \left(\tau_{y,a_s(m)}^s \right)^{-1}}{\prod_{i=1}^4 \left(\frac{\tau_{i,\tilde{m}(s)}^s p_i^x}{\theta_i^{a_s(\tilde{m}(s))}} \right)^{\theta_i^{a_s(\tilde{m}(s))}} \left(\tau_{y,a_s(\tilde{m}(s))}^s \right)^{-1}}, \tag{33}$$

where we normalize $z_{\tilde{m}(s)}^s = 1$.

The only parameters left to calibrate are the endowments of the inputs.

$$p_i^x x_i = \sum_{s=1}^{S} \sum_{m=1}^{M^s} \frac{\tau_{y,a_s(m)}^s}{\tau_{i,m}^s} \theta_i^{a_s(m)} \frac{\rho^s - 1}{\rho^s} p_m^s y_m^s$$

We remove all distortions from both of the calibrated economies. Table VIII reports the changes in industry-level and aggregate output. We find that if we only consider one technology then the gains in aggregate output are 41.5 percent. On the other hand, if we consider multiple technologies then the increase in output is 19.7 percent. Thus, we find significant declines in output gains from the elimination of all misallocation in the economy if we allow for multiple production technologies.

Table VIII also reports the changes in industry-level and aggregate TFP. To calculate TFP, we first construct a representative production function for each technology. Output for technology h in industry s (industry notation is still supressed) is

$$Y_h = \left(\sum_{m=1}^{M} \mathbb{I}_{h=a(m)} \left(y_m \right)^{\rho} \right)^{1/\rho}. \tag{34}$$

Thus, the TFP for technology h is

$$z_h = \frac{Y_h}{\left(x_{k,h} \right)^{\theta_k^h} \left(x_{sl,h} \right)^{\theta_{sl}^h} \left(x_{ul,h} \right)^{\theta_{ul}^h} \left(x_{m,h} \right)^{\theta_x^h}}, \tag{35}$$

where $x_{k,h}$, $x_{sl,h}$, $x_{ul,h}$, and $x_{m,h}$ is the total amount of capital, skilled labor, unskilled labor, and intermediate inputs used by firms utilizing technology h. To calculate changes in industry-level TFP, we find the percentage change in TFP for each technology after the removal of distortions.

We then weight these percentage changes with the share of gross output accounted for by firms that use that technology in the calibrated economy.

We find that the aggregate gains in TFP are 13.6 percent under multiple technologies and 34.7 percent if we only allow for one technology. As before, we find significant declines in gains from eliminating all misallocation in the economy. As expected, we find consistent TFP increases across all industries. This finding is in contrast to changes in output, in which we find that some industries see declines in output after the removal of distortions. This result highlights the importance of the re-allocation of inputs across sectors.

5 Conclusion

In this paper, we introduced a methodology for identifying different production technologies (Cobb-Douglas factor intensities in the production function) used by firms within an industry using clustering analysis. We show through simulations that our methodology performs with high accuracy when input distortions across firms appear lognormally distributed, which is consistent with what we observe in the data. Being able to identify multiple production technologies is critical when studying misallocation, since the way misallocation is typically identified in the data is by using variations in observed input cost shares across firms. Using plant level data for Chilean manufacturing firms, we find robust evidence that these cost shares differ across firms due to both distortions and different production technologies. In the case of the furniture industry in Chile, this causes gains from reducing misallocation to be overstated according to the standard, single-production technology analysis. More importantly, clumping firms together masks significant variation across production technologies in the degree of misallocation, as we find evidence that smaller firms tend to use different production technologies than large firms and be subject to greater magnitudes of distortions. Our results suggest that accounting for differences in production technologies may be important for understanding the misallocation gap between developed and developing countries, although more work is needed to know whether the magnitude of the gap will increase or decrease once these adjustments are made.

Table VI

Decline (%) in the Standard Deviation of Input Distortions

One vs. Multiple Technologies

ISIC	Capital	Skilled Labor	Unskilled Labor	Intermediate Inputs
1511	0.0	0.0	0.0	0.0
1512	29.5	7.6	9.0	0.4
1513	15.2	7.8	23.8	0.7
1520	18.2	24.0	17.8	0.1
1531	16.1	14.6	23.1	3.1
1541	39.2	2.5	0.8	0.9
1552	25.2	18.1	22.4	2.8
1711	8.2	24.0	39.7	3.7
1810	18.0	5.0	21.6	2.3
1920	39.5	31.6	9.5	6.4
2010	25.3	5.1	21.0	3.4
2102	25.2	24.7	17.2	7.4
2221	0.5	7.3	48.4	0.1
2411	20.8	9.5	20.5	1.3
2424	7.0	14.9	29.7	0.5
2429	30.0	11.1	8.7	4.6
2520	21.3	12.8	8.8	1.6
2695	31.7	15.4	17.5	3.7
2720	31.8	28.8	24.3	1.0
2811	29.0	4.1	12.7	2.3
2899	15.6	1.8	22.9	1.4
3610	32.4	3.1	13.3	72.0
mean	21.8	12.4	18.8	5.4

Notes: Table VI shows the percent decline in the standard deviation of the inputs distortions when going from one to multiple technologies. Column 1 reports the industry code. Columns 2-5 show the percent decline in the standard deviation of the logged input distortions for capital, skilled labor, unskilled labor, and intermediate inputs respectively. The percent decline of the standard deviation for input i is calculated as $-(\sigma_{i,multi}/\sigma_{i,onetech}-1)*100$, where $\sigma_{i,multi}$ is the standard deviation of the distortion for input i with multiple technologies and $\sigma_{i,onetech}$ is the same statistic for the case in which we only allow for one technology.

Table VII

Recovered Production Technologies

Manufacturing of Non-Ferrous and Precious Metals (ISIC 2720)

Case	Technology	θ_{sl}	θ_{ul}	θ_k	θ_x	Count
Multiple technologies	1	0.03	0.02	0.01	0.94	23
Multiple technologies	2	0.10	0.08	0.05	0.77	34
Single technology	1	0.08	0.06	0.04	0.82	57

Notes: Table VII shows the recovered factor intensities for the manufacturing of non-ferrous and precious metals (ISIC 2720). Column 1 indicates whether we report the case in which we impose a single technology or allow for multiple technologies. Column 2 lists the technology number, which is useful if there is more than one technology in the industry. Columns 3-6 report the factor intensities of skilled labor, unskilled labor, capital, and intermediate inputs respectively. Column 7 lists the number of plants that use the technology.

Table VIII

Gains in Output and TFP after Removing all Distortions (%)

One vs. Multiple Technologies

	Output			TFP		
ISIC	Multiple Technologies	One Technology	Percent Change	Multiple Technologies	One Technology	Percent Change
1511	21.6	19.0	13.7	17.3	17.3	0.0
1512	20.4	26.6	-23.5	19.0	29.8	-36.2
1513	15.0	17.3	-13.4	12.2	19.5	-37.2
1520	21.5	23.2	-7.6	6.1	13.7	-55.1
1531	25.6	23.4	9.6	6.2	11.5	-46.1
1541	0.4	-4.2	-109.9	17.5	23.3	-24.9
1552	42.4	84.8	-50.0	49.4	107.8	-54.1
1711	17.9	27.6	-35.2	14.6	29.3	-50.2
1810	17.2	26.4	-34.8	12.6	31.3	-59.7
1920	8.7	4.1	113.2	12.8	18.2	-29.9
2010	21.7	24.6	-11.6	16.4	25.3	-35.1
2102	11.5	12.7	-9.9	8.0	10.2	-21.8
2221	14.9	21.7	-31.5	9.7	27.3	-64.4
2411	14.6	66.5	-78.1	9.9	37.7	-73.7
2424	-0.9	2.7	-132.7	9.4	19.1	-50.8
2429	21.5	21.9	-1.8	14.4	19.5	-26.4
2520	21.9	21.3	2.8	15.0	18.8	-20.6
2695	9.7	14.2	-32.0	15.3	23.0	-33.5
2720	25.4	48.5	-47.6	14.1	44.3	-68.2
2811	24.8	24.5	0.9	15.8	23.6	-33.0
2899	43.2	56.7	-23.9	16.5	37.8	-56.2
3610	9.1	9.5	-4.0	12.4	19.3	-35.9
Aggregate	19.7	41.5	-52.7	13.6	34.7	-60.6

Notes: Table VIII reports the percent increase in industry-level and aggregate output from eliminating all distortions. Column 1 reports the industry code. Column 2 lists the calculated increase in output upon removing all distortions when we allow for multiple technologies. Column 3 reports the same statistic except that we only allow for one technology. Column 4 shows the percent change in output increases when we allow for multiple technologies relative to one technology. Column 5, 6, and 7 report the same statistics as columns 2, 3, and 4 for increases in TFP.

References

Arthur, D., and S. Vassilvitskii (2007): "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Bergoeing, R., and A. Repetto (2006): "Micro Efficiency and Aggregate Growth in Chile," *Latin American Journal of Economics*, 43(127), 169–192.

Calinski, T., and J. Harabasz (1974): "A dendrite method for cluster analysis," *Communications in Statistics*, 3(1), 1–27.

Hsieh, C.-T., and P. Klenow (2009): "Misallocation and Manufacuring TFP in China and India," *Quaterly Journal of Economics*, 124(4), 1403–1448.

Levinsohn, J. (1999): "Employment responses to international liberalization in Chile," *Journal of International Economics*, 47(2), 321–344.

Levinsohn, J., and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70(2), 317–341.

Pavcnik, N. (2002): "Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants," *The Review of Economic Studies*, 69(1), 245–276.

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

Steinley, D., and M. J. Brusco (2011): "Evaluating mixture modeling for clustering: Recommendations and cautions," *Psychological Methods*, 16(1), 63–79.

Young, A. (1995): "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experience," *The Quarterly Journal of Economics*, 110(3), 641–680.