Misallocation in the Presence of Multiple Production Technologies*

Jose Asturias
Georgetown University Qatar

Jack Rossbach
Georgetown University Qatar

October 12, 2017

Abstract

Whereas firms may use different production technologies to produce the same or similar goods, this information is typically not available to researchers. We develop a methodology to identify when multiple production technologies are present in an industry and to classify firms by which production technology they use. Our strategy works by applying cluster analysis to detect groupings in the distributions of factor expenditure shares. We apply our methodology to Chilean plant-level data and find that multiple production technologies operate in the majority of industries. We show that accounting for the presence of multiple production technologies reduces the predicted gains in output from eliminating misallocation by approximately 30 percent.

JEL classifications: D24, O11, O47, O14

Keywords: Cluster Analysis, Misallocation

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

1 Introduction

It is commonly understood that many goods can be produced using a variety of production technologies, each of which may require different combinations of factor inputs to produce the same or similar output. For example, 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 using a fully mechanized assembly line. Multiple production technologies can also operate in industries where products are relatively homogeneous, for example, in the production of copper concentrates. Whereas 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, it is therefore seen as difficult or impossible to identify the various technologies used 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 multiple industries simultaneously, however, we are not aware of any existing methodologies for identifying whether multiple production technologies are used within an industry and which technologies are used by which firms.

We develop a methodology to identify the production technologies operating within an industry and to identify which technology is used by each firm. The methodology uses observed variation in cost shares (expenditures on each factor input divided by revenues) across firms. Importantly, our methodology does not require prior knowledge on the number or types of production technologies operating within an industry. Our methodology is based on cluster analysis, which works by identifying groups of observations that share similar characteristics — in our case, cost shares. Cluster analysis is common in applications where pattern recognition is important, for example, image processing and population inference based on genotype data.[1] Cluster analysis allows us to separate idiosyncratic variation in cost shares across firms — for example, variation due to idiosyncratic distortions — from systematic variation, which we argue arises when firms use different production technologies. We view the results of the cluster analysis through the lens of a structural model, which is an innovation that distinguishes our methodology from previous applications of cluster analysis in economics.[2] In particular, our model allows us to show how the cluster analysis should

[1]See Jain (2010) for a survey of the use of clustering in pattern recognition. For examples in the field of image processing, see Park, Yun, and Lee (1998) and Ray and Turi (1999); for examples in the field of population inference see Jakobsson and Rosenberg (2007) and Pritchard, Stephens, and Donnelly (2000).

[2]An early and influential application of cluster analysis in economics was recovering descriptive groups among qualitative marketing and business data. For example, Gartner (1990) surveys researchers and business owners on their personal definition of entrepreneurship and uses cluster analysis to classify responses into two distinct views of what entrepreneurship entails. Kotey and Meredith (1997) survey small furniture manufacturers about the relationship between owners and managers and use cluster analysis to group respondents into three relationship types. Leuz, Nanda, and Wysocki (2003) conduct a descriptive cluster analysis in which they group countries using

be applied to the data, how to interpret the results, and under what conditions our methodology will successfully identify heterogeneous production technologies, which we verify using simulated data.

Our methodology for identifying production technologies is conservative in the sense that we only identify production technologies that are sufficiently distinct in terms of their relative factor input intensities compared to the idiosyncratic variation in cost shares across firms. This means, for example, whereas 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 distinguish between two similar labor-intensive technologies. In this regard, what we call a production technology might be considered as a set of closely related production technologies, which we group together. In the future, and with additional data on firms, such as intermediate input usage, our methodology could be used to distinguish more closely related production technologies.

We apply our methodology to Chilean plant-level manufacturing data and find strong evidence of multiple production technologies in the majority of industries. We find that 14 of the 23 industries considered have more than one production technology: 7 industries have 2 technologies, 5 industries have 3 technologies, and 2 industries have 4 technologies. We find that when operating under the assumption of a single production technology, the resulting factor intensities are effectively a weighted average of the factor intensities for each of the production technologies operating within an industry. We explore our results and find that the production technologies recovered by our methodology match many intuitive features of the data. For example, we find that smaller firms are more likely to use labor-intensive technologies. This is the case even though information on firm size is not used as part of our methodology for recovering production technologies.

After establishing the presence of multiple production technologies in a majority of the industries that we consider, we evaluate the quantitative importance 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 across firms can be a major factor in explaining the gaps in output and total factor productivity (TFP) between rich and poor countries. An influential paper in this field is that of Hsieh and Klenow (2009), who determine the TFP losses in China and India

data on legal and institutional characteristics. More recent applications of cluster analysis include recovering trend patterns in quantitative data. Crone (2005) uses cluster analysis to group states into economic regions based on similarity in business cycle characteristics, Levy-Yeyati and Sturzenegger (2005) classify countries into exchange rate regimes using data on exchange rates and international reserves, and Humphries (2017) classifies individuals according to their life-cycle employment profiles. There are two major differences between our paper and previous applications of cluster analysis. First, we develop an economic model that tells us which variables should be included in the cluster analysis, how the cluster analysis should be applied, and how the output can be tied to specific structural parameters in our model. Second, we develop a new methodology for statistically determining the number of groups in our cluster analysis, whereas previous applications have relied on heurstic methods for determining the number of groups.

accounted for by misallocation. They infer misallocation using a model of monopolistic competition and backing out the wedges that rationalize the first-order conditions of firms along with the observed choice of inputs. These wedges result in a dispersion in the marginal revenue product of inputs across firms in the model.[3] Estimating the efficient allocation of inputs requires researchers to specify a production function for each firm in the economy. The standard method for doing so is to assume that production functions are identical for all firms within each industry, except for a Hicks-neutral productivity parameter that varies across firms. This setup implies that all firms should have the same cost shares for each input; however, in the data we observe significant variation in cost shares even for firms within narrowly defined products and industries. Misallocation is then measured by assuming that variation in observed cost shares arises from unobserved distortions on the factor input prices that firms face. Removing these distortions leads to reallocation of factor inputs across firms and thus increases aggregate output and TFP. Applications of this methodology are found in a wide variety of settings, including the studying of misallocation during times of crisis (Oberfield (2013), Sandleris and Wright (2014)), and misallocation in southern Europe (Dias, Marques, and Richmond (2016), García-Santana, Moral-Benito, Pijoan-Mas, and Ramos (2016), Gopinath, Kalemli-Ozcan, Karabarbounis, and Villegas-Sanchez (2015)), misallocation in the United States in the 1800s (Ziebarth (2013)) and across Latin American countries (Busso, Madrigal, and Pagés (2013)).

Our methodology is important for measuring misallocation as we are able to loosen the standard assumption that all firms within an industry share a common production technology. If firms with different production technologies are assumed to have the same production function, we would incorrectly classify all differences in cost shares as misallocation, instead of taking into account that some of the differences arise from differing production technologies. We find that accounting for the presence of multiple production technologies results in significant reductions in the dispersion of marginal revenue products for each input. For example, the standard deviation of model-implied distortions for capital declines by an average of 31.0 percent across industries. We find that dispersion in the marginal revenue products for skilled and unskilled labor is strongly correlated with size, which Restuccia and Rogerson (2008) show amplifies the misallocation that results from the dispersion and leads to larger losses in overall output and TFP. The strength of the relationship is significantly overstated, however, when a single production function is assumed, as we find that large firms are less likely to use labor-intensive technologies. In contrast, we find that the marginal revenue products of capital and intermediate inputs are uncorrelated with firm size even under the assumption of a single production technology.

We conduct a quantitative misallocation exercise similar to Hsieh and Klenow (2009) where we

[3] The use of wedges to identify misallocation is usually referred to as the "indirect" approach and was also pioneered by Guner, Ventura, and Xu (2008) and Restuccia and Rogerson (2008). A similar set of papers back out the distribution of distortions that match the firm size distribution across countries, including Alfaro, Charlton, and Kanczuk (2009), and Bartelsman, Haltiwanger, and Scarpetta (2013).

estimate the counterfactual gains in total output from the elimination of distortions on factor input prices. We conduct this exercise under the standard assumption that all firms in each industry share a single production function and then again after using our methodology to identify the heterogeneous production technologies within industries. We find that accounting for the presence of multiple production technologies reduces the predicted gains in total output by approximately 30 percent. This implies that accounting for multiple production technologies is important for understanding the potential gains from reducing misallocation and therefore, potentially, the fraction of variation in GDP per capita across countries that can be explained by differences in misallocation.

Our paper is the first to show that a significant amount of traditionally measured misallocation can be explained by differences in production technologies across firms. Our findings are useful in helping to answer the question in the misallocation literature of what fraction of measured misallocation reflects true misallocation as opposed to other sources of variation, for example, model misspecification and data mismeasurement. Hopenhayn (2014) discusses possibilities such as measurement error, the degree of decreasing returns to scale, and capital adjustment costs. Asker, Collard-Wexler, and De Loecker (2014) find that capital adjustment costs can potentially account for a large portion of the wedges on capital and therefore the observed dispersion in the marginal revenue product of capital. Other papers that explore sources of mismeasurement include those by Li (2015), Schelkle (2016), and Bils, Klenow, and Ruane (2017). We expect that our findings will be complementary to most of the alternative explanations for measured misallocation in the literature, such as adjustment costs and measurement error. This is because, for example, adjustment costs lead to dispersion in cost shares around a single point, whereas firms using different production technologies leads to gaps in the distribution and multiple centers around which there is dispersion in cost shares. Therefore, we are targeting a different type of dispersion compared to these other papers, although we would not be surprised if adjustment costs and measurement error explain a significant amount of the misallocation that remains after controlling for multiple production technologies. Likewise, multiple production technologies can explain some features of the data that are not easily explained by these alternatives. For example, the misallocation of unskilled and skilled labor, where capital adjustment costs are not relevant, is a significant source of misallocation in our dataset. Furthermore, the distortions for those inputs are correlated with firm size. Our finding that larger firms use less labor-intensive technologies provides a natural explanation for this correlation, whereas alternatives such as measurement error may struggle in this regard.

2 Model

In this section, we present the theoretical framework we use to motivate our methodology for identifying production technologies. We first lay out our framework for an industry featuring multiple production technologies in Section 2.1 without specifying which features of the economy are observable in the data. In Section 2.2, we then motivate which parameters of the economy we treat as observable in the data and describe the intuition behind our strategy for identifying parameters that are not directly observable.

2.1 An Industry with Multiple Production Technologies

The framework we use is a standard model of monopolistic competition with heterogeneous firms. Our methodology for identifying production technologies within an industry will only require data for that industry; therefore, we focus this section on firms within a single industry. We consider a multi-industry extension of the model when we perform counterfactual exercises in Section 4.4.

The industry contains M firms, where each firm produces a differentiated product according to a constant returns to scale Cobb-Douglas technology. Our framework is similar to that of Hsieh and Klenow (2009), except 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 firm m's output, z_m is its productivity, x_m^i is its quantity of input i, and θ_h^i is the factor intensity for input i by firms using technology h. We follow the standard assumption that production technologies exhibit constant returns to scale; therefore, $\sum_{i=1}^{I} \theta_h^i = 1$ for all technologies h. We assume that a firm's technology is exogenously given rather than an endogenous choice.

The industry has a perfectly competitive bundler that combines the differentiated outputs into an industry output according to a constant elasticity of substitution 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

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

where E is total expenditures on the industry output, which are constant, 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 price of factor input i, τ_m^i is the idiosyncratic distortion that firm m faces for input i, and τ_h is a distortion that is common to all firms employing production technology h and which does not depend on the specific input. This distortion can be thought of as a tax or subsidy to the total costs that firms face and depends only on their production technologies. Despite referring to the τ's as distortions, our methodology does not rely on this interpretation until we perform counterfactuals in Section 4. The essential point here is that there is a wedge between the factor input expenditures that we observe in the data and the firm's true factor input expenditures. This wedge could be due to unobserved distortions, as we assume, or due to other sources such as measurement error.

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

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

This problem has the solution that firms charge a constant markup, equal to $1/\rho$, over their marginal cost,

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

where the marginal cost for firm m depends on the firm's productivity, the firm's distortions, and the firm's production technology. This expression also highlights that the impact that distortions have on a firm's marginal cost depends on the firm's production technology.

The presence of these distortions affects the choices of firms when choosing inputs and leads to dispersion in the marginal revenue products of inputs across firms. In this case, the marginal revenue product (MRP) of input i for firm m using production technology h is given by

$$MRP_{h,m}^i = \tau_h \tau_m^i p^i. \tag{7}$$

Thus, if a firm faces a large distortion, τ_m^i, on input i, then the firm will use less of that input, and that input will have a correspondingly high marginal revenue product. When we perform counterfactuals in Section 4, we interpret these distortions as causing firms to use inefficient amounts of each input, which leads to misallocation.

Cost minimization implies that the share of a firm's total expenditures, including distortions, spent on a given input will be equal to the firm's factor intensity for that input:

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

This equation highlights that all firms with the same production technology will have the same expenditure shares on each input when the unobserved distortions are included.

Finally, we can show that the following condition holds:

$$\frac{1}{\rho} = \frac{p_m y_m}{\tau_h \sum_{j=1}^{I} \tau_m^j p^j x_m^j}, \tag{9}$$

which relates markups to revenues and factor expenditures with distortions. This condition will be useful in the following section for identifying model parameters.

2.2 Connecting the Model to Data

Up to this section, we have not discussed which features of the model are typically available in the data and which are not. The motivation for this paper, however, is that many of the parameters in Section 2.1 are not available in standard data. This unavailability is the reason most papers are forced to assume that only a single production technology is used in each industry. For example, if factor intensities were directly observable, then we would instantly observe 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 directly recover the factor intensities and therefore the production techniques.

We are interested in cases where information on the primary model parameters are not directly observable in the data. Furthermore, whereas some datasets break down values into prices and quantities, many datasets contain only expenditures. Therefore, we assume that we are only able to observe expenditures, excluding distortions, $p^i x^i_m$, for each input i and revenues, $p_m y_m$, for each firm m in the industry. Note that our observed expenditures on inputs do not include the unobserved idiosyncratic distortions or the technology-specific distortions. In addition, we assume that we are not able to observe markups or factor intensities, nor are we able to observe the production technology used by each firm.

To derive expressions that can be used to recover the unobservable parameters, we first define the revenue cost share for input i as

$$\tilde{\theta}^i_m \equiv \frac{p^i x^i_m}{p_m y_m}. \tag{10}$$

These cost shares do not include distortions and are therefore directly observable in the data. We can then combine equation 8, which relates input expenditures with distortions and factor intensities, with equation 9, which relates input expenditures to the markup over marginal cost, to derive the following expression for the cost share for firm m for input i:

$$\tilde{\theta}^i_m = \frac{\rho \theta^i_h}{\tau_h \tau^i_m}. \tag{11}$$

This equation relates each firm's observed expenditures on input i to its technology-specific distortion, τ_h, its idiosyncratic distortion for that input, τ^i_m, the markup, ρ, and its factor intensity for that input, θ^i_h. This expression will be the main equation we use throughout the paper to identify the unobservable parameters and to identify the production technology used by each firm. The key benefit of this equation is that the cost share for a given input does not depend on the factor intensities or distortions for other inputs. In contrast to dividing by revenue, if we were to instead use observed input expenditures as a fraction of total observed expenditures, $p^i x^i_m / \sum_{j=1}^{I} p^j x^j_m$, then the cost share for each input would depend on all of the idiosyncratic distortions simultaneously.

Our strategy for recovering production technologies and other unobserved parameters is to use variation in cost shares to pin down factor intensities and distortions. The key insight is that the variables in equation 11 all vary differently across firms, inputs, and production technologies. For example, if we compare two firms that employ the same production technology, their cost share will vary solely because of idiosyncratic distortions and not because of the other parameters. This implies that we need to determine which production technology is used by each firm. In the following section, we describe our methodology for recovering production technologies. In Section 3.4, we then show how to use this information to recover factor intensities, markups, and distortions.

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 is based on cluster analysis, which is a set of processes used to group observations so that they are similar within groups. Our application of cluster analysis will be based on the expression for cost shares derived in the previous section and on the intuition that, in an industry with multiple production technologies, we should expect gaps in the dispersion of cost shares to arise between sets of firms with different production technologies.

We present our methodology in several steps. First, in a preliminary step, we transform the cost shares to prepare for the cluster analysis. Second, we partition the set of firms in each industry using the k-means clustering algorithm for an arbitrary number of groups. Third, after applying the clustering algorithm to a range of possible groups, we use those results to recover the number of production technologies within each industry. Fourth, we show how to use our clustering results to identify the underlying parameters of the model: markups, factor intensities, and distortions. In the appendix, we verify that our methodology works in that it is able to successfully recover the number of production technologies in each industry and the assignment of firms to these production technologies using simulations based on synthetic data.

3.1 Transforming Data

Our methodology relies on exploiting the structure of equation 11.[4] Before applying cluster analysis, we take the log of the equation to get

$$\log \tilde{\theta}^i_m = \log \rho + \log \theta^i_h - \log \tau_h - \log \tau^i_m. \tag{12}$$

[4]Note that equation 11 is similar to equation 18 in Hsieh and Klenow (2009), which they use to identify the distortions in their model. In general, most papers in the misallocation literature rely on identifying distortions based on variation in expenditure shares. This means our framework is relatively standard and our methodology for identifying production technologies can be easily incorporated into frameworks in the misallocation literature with alternative specifications.

This transformation is useful because it linearizes the relationship between the cost shares. Furthermore, a common assumption in the literature is that idiosyncratic distortions follow a lognormal distribution. This assumption appears to be satisfied when we apply our methodology to Chilean plant-level data in Section 4. In addition, although the clustering algorithm we apply is non-parametric and does not rely on the assumption of lognormality, a large literature has shown that the k-means clustering algorithm works particularly well for clustering Gaussian-distributed data.[5] Therefore, transforming the data in the above way helps to ensure the accuracy of our results.

3.2 Partitioning Firms by Production Technology

In order to recover the number of production technologies, we follow a two-step strategy. First, in this subsection, we show how to optimally partition the set of firms in order to determine which production technology each firm uses under the assumption that there are $\mathcal{H}$ production technologies in the industry. Second, in the next subsection, we show how to use the sum of squared errors (SSE) from the optimal partition for a range of possible production technologies, $\mathcal{H} = 1, ..., \mathcal{H}_{\max}$, where $\mathcal{H}_{\max}$ is the maximum number of technologies we consider, to identify the actual number of production technologies, H, in an industry.

Under the assumption that there are $\mathcal{H}$ production technologies in an industry, our goal is to partition the set of firms into $\mathcal{H}$ groups where each firm is assigned to a single group, $h \in \{1, ..., \mathcal{H}\}$, and each group has at least one firm assigned to it. For each grouping, our optimal partition minimizes the SSE between the logged cost shares and the mean logged cost share for each firm's group. In particular, the optimal partition achieves

$$SSE_{\mathcal{H}} \equiv \min_{\{\hbar(m)\}} \sum_{h=1}^{\mathcal{H}} \sum_{m=1}^{M} \sum_{i=1}^{I} \mathbb{I}_{\hbar(m)=h} \left(\log \tilde{\theta}_m^i - \xi_h^i\right)^2, \tag{13}$$

where $\hbar(m)$ is the mapping that returns the group that firm m is assigned to, $\mathbb{I}_{\hbar(m)=h}$ is an indicator function that takes on the value of one if firm m is assigned to group h and zero otherwise, and ξ_h^i is the mean value of the logged cost shares for input i for firms assigned to group h.

Our motivation for seeking the partition that minimizes the SSE is that, under two assumptions, this is the partition that will assign each firm to its actual production technology. The first assumption that needs to be satisfied is that $\mathcal{H}$ is the true number of production technologies operating within the industry. The second assumption is that each firm's cost share is closest to the average cost share for its production technology. In particular, we will assume that the Euclidean distance between each firm's cost shares are closest to the undistorted cost shares for its actual production technology. Formally, this condition is equivalent to

[5] Jain (2010) has an overview on the use and development of the k-means algorithm and points out that the algorithm does well in finding spherical clusters in the data, which is consistent with a Gaussian distribution.

$$\sum_i \left(\log \tau_h + \log \tau_m^i\right)^2 < \sum_i \left[\log \theta_{h'}^i - \left(\log \theta_h^i - \log \tau_h - \log \tau_m^i\right)\right]^2, \ \forall h' \in \{1, ..., \mathcal{H}\}, h' \neq h. \quad (14)$$

The practical importance of this assumption is that, on average, our estimated distortions will be as small as possible in their magnitude.

In general, it is impossible to find the globally optimal partition that minimizes the SSE to solve the problem in equation 13 because of the high dimensionality of the problem, which belongs to the class of NP-hard problems. Even with only two technologies and 50 firms, there are 2^{49} possible groupings, which highlights the infeasibility of the brute force method of evaluating all possible combinations to find the partition that minimizes the SSE. To tackle this problem, we apply the heuristic k-means++ clustering algorithm developed by Arthur and Vassilvitskii (2007) to find a grouping that achieves a local minimum of SSE. The k-means++ algorithm is initialized by selecting $\mathcal{H}$ centers randomly from the set of observations. The first center is chosen uniformly at random, and subsequent centers are chosen with probability inversely proportional to their Euclidean distance from the closest previously chosen center (in the standard k-means algorithm, all $\mathcal{H}$ centers are chosen uniformly at random). This initialization strategy ensures that the starting points tend to be far away from each other and improves the speed and accuracy of the algorithm compared to selecting all centers with uniform probability. Following initialization, the algorithm iterates between the following two steps until convergence:

- Assignment step: Assign each observation to the closest center:

$$\hbar(m) = \underset{h \in \{1,...,\mathcal{H}\}}{\arg\min} \sum_{i=1}^{I} \left(\log \tilde{\theta}_m^i - \xi_h^i\right)^2. \quad (15)$$

- Recentering step: After all observations have been assigned, recompute centers as the component-wise mean values for observations in each group:

$$\xi_h^i = \frac{1}{\sum_{m=1}^M \mathbb{I}_{\hbar(m)=h}} \sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \log \tilde{\theta}_m^i. \quad (16)$$

After convergence is achieved, meaning that the centers remain the same and no observations switch group assignment, the algorithm will yield a partition where each firm is assigned to one of $\mathcal{H}$ groups. If $\mathcal{H}$ is equal to the number of production technologies in the industry — determining that this is the focus of the next step of our methodology — and the partition is the globally optimal one, then this will correctly separate firms by their production technologies. To avoid selecting a locally optimal partition far from the globally optimal partition, we reinitialize the algorithm many times with different randomly drawn starting points and then select the partition with the lowest SSE.

3.3 Determining the Number of Production Technologies

Our goal is to recover the total number of production technologies, H, used by firms within the industry. After determining the number of production technologies, we can then partition the set of firms following Section 3.2 to recover the production technology used by each individual firm. Our strategy for recovering H will again be based on the intuition that there should be gaps in the distributions of cost shares between firms with different production technologies. Effectively, our clustering algorithm in Section 3.2 ensures that firms are similar within groups, and our methodology in this section will ensure that firms are dissimilar across groups and therefore that firms that share a common production technology are not arbitrarily split into separate groups.

To determine the number of production technologies in an industry, we cluster our data using the clustering algorithm in Section 3.2 with $\mathcal{H} = 1, ..., \mathcal{H}_{\max}$ groups, where again $\mathcal{H}_{\max}$ is the maximum number of groups we consider. Our strategy relies on computing the SSE of the optimal partition, $SSE_{\mathcal{H}}$, for each grouping, $\mathcal{H}$, and then comparing the reduction in SSE of going from $\mathcal{H} - 1$ to $\mathcal{H}$ groups.

We want to be conservative in our estimate of the number of production technologies in order to avoid claiming the existence of multiple production technologies when there is only a single one. Therefore, we evaluate the presence of additional production technologies in a sequential matter. This means we first test whether there is evidence of two production technologies in the industry versus the null hypothesis that there is only one. Increasing the number of groups in the clustering algorithm always leads to a decrease in the SSE (up to the number of observations, at which point the SSE $= 0$). Therefore, we need to test whether the reduction in the SSE from clustering the data with two groups instead of one is greater than the reduction we would expect if we knew there was only a single production technology. Conditional on finding a sufficient reduction in the SSE to reject the null hypothesis that there is only a single production technology, we continue the process. Specifically, we examine whether the reduction in the SSE from clustering the data with three groups instead of two is greater than the reduction we would expect if there were only two production technologies, and so on.

In order to compute the expected reduction in the SSE conditional on having a given number of production technologies, we require an assumption on the data-generating process.[6] In particular, we assume that the distortions on factor inputs, τ_m^i, follow a multivariate lognormal distribution. This assumption is not particularly restrictive, as our methodology still allows both for distortions to be correlated across factor inputs — capturing that firms with a high capital cost distortion may be more likely to have a high labor cost distortion — and for the distribution of distortions to differ

[6]We can avoid assumptions on the data-generating process if we instead use the widely used C-H index from Calinski and Harabasz (1974), which is a heuristic measure used to determine the number of production technologies. The C-H index cannot be used to determine whether there is only one production technology, however, the C-H index yields identical results when we apply our methodology to Chilean plant-level data in Section 4 for industries we identify as having with more than one technology.

across production technologies within an industry. Additionally, the assumption of lognormality appears to be well supported by our data, which we show using kernel density plots in Section 4.

We compute the expected reduction in the SSE from clustering with 2 groups conditional on having only 1 production technology in three steps. First, we fit a single multivariate lognormal distribution to the data. Second, we generate a large number of synthetic datasets with the same number of observations as the original dataset, where each observation is drawn independently from the fitted multivariate lognormal distribution. Third, we apply our clustering algorithm from Section 3.2 individually to each synthetic dataset and calculate

$$R_{2|1}^{\text{syn}} = \frac{SSE_2^{\text{syn}}}{SSE_1^{\text{syn}}}, \tag{17}$$

where SSE_2^{syn} and SSE_1^{syn} are calculated using equation 13. $R_{2|1}^{\text{syn}}$ is the ratio of the SSE from running our clustering algorithm with 2 groups versus 1 group for the synthetic dataset, where, by construction, the synthetic dataset has only 1 true production technology. A lower $R_{2|1}^{\text{syn}}$ therefore indicates a greater reduction in the SSE from clustering with 2 groups versus 1 group.

After computing $R_{2|1}^{\text{syn}}$ for each synthetic dataset, we define $R_{2|1}^{\hat{\text{syn}}}$ to be the 20th percentile of $R_{2|1}^{\text{syn}}$ values, so that 80 percent of the synthetic datasets have a reduction in the SSE that is smaller than $R_{2|1}^{\hat{\text{syn}}}$ (we discuss the intuition behind using the 20th percentile in the next paragraph). We use $R_{2|1}^{\hat{\text{syn}}}$ as our cutoff test statistic for determining whether the actual data contain a single production technology or multiple production technologies. In particular, we compute $R_{2|1} = SSE_2/SSE_1$ for our original data and then run the following comparison:

$$\text{Conclude 2 or more production technologies} \iff R_{2|1} < R_{2|1}^{\hat{\text{syn}}}. \tag{18}$$

This is similar to performing a hypothesis test where the null hypothesis is that there is a single production technology, and we reject the null hypothesis if $R_{2|1} < R_{2|1}^{\hat{\text{syn}}}$. This test is based on the intuition that we are unlikely to get a greater reduction in the SSE in the data without having detectable "gaps" in the distribution of cost shares.

We choose the 20th percentile because it leads to robust results when applied to our data in Section 4. In particular, our results are not affected if we switch to using the 15th or 25th percentile. This percentile is not the same as the significance level of our test (in 18, the significance level is the probability that we conclude there are two or more technologies when the true number of technologies is one). The percentile we use would only be the significance level if the true data-generating process is the same as the one we used to generate the synthetic datasets. The true data-generating process, however, may differ significantly from our fitted distribution. This could be, for example, because the synthetic datasets are generated assuming a single production technology, whereas the original data contain more than one production technology. In simulations, we find a significance level much smaller than 5 percent when we use the 20th percentile of $R_{2|1}$ for our cutoff. In general, the cost of using lower percentiles, such as the 5th percentile, is that we will suffer from increased type II error without having much impact on our type I error. On the other hand, if we use a cutoff much higher than the 20th percentile, we begin to start suffering from increased type I error, which goes against our goal of being conservative in our estimates for the number of production technologies in each industry.

Conditional on concluding there is evidence for two or more production technologies, we repeat our exercise to evaluate whether there is evidence for three or more production technologies, and we continue this process sequentially. Our general methodology for concluding there are $\mathcal{H}$ or more production technologies — conditional on already concluding there are $\mathcal{H}-1$ or more production technologies — requires an estimate of the expected reduction in the SSE from clustering with $\mathcal{H}$ groups if there are only $\mathcal{H}-1$ production technologies. We calculate this in four steps. First, we cluster the data according to Section 3.2 with $\mathcal{H}-1$ groups. Second, we fit a multivariate lognormal distribution independently to each of the $\mathcal{H}-1$ groups in our clustered data. Third, we generate a large number of synthetic datasets, where each synthetic dataset is constructed by taking independent draws from each of the $\mathcal{H}-1$ multivariate lognormal distributions, such that the number of observations drawn from each distribution is equal to the number of observations in the group to which we fitted the distribution in the original data. For example, if we use $\mathcal{H}-1=2$ groups in our original data, and after clustering, group 1 contains 15 observations, whereas group 2 contains 20 observations, then our synthetic datasets will contain 15 observations drawn from the multivariate lognormal distribution corresponding to group 1 and 20 observations drawn from the multivariate lognormal distribution corresponding to group 2. Fourth, we apply our clustering algorithm from Section 3.2 to each synthetic dataset individually and calculate $R^{\text{syn}}_{\mathcal{H}|\mathcal{H}-1}$, which we do using an equation analogous to 17.

We set $R^{\hat{\text{syn}}}_{\mathcal{H}|\mathcal{H}-1}$ as the 20th percentile of $R^{\text{syn}}_{\mathcal{H}|\mathcal{H}-1}$ datasets, and we conclude there is evidence for $\mathcal{H}$ or more production technologies if the reduction in the SSE in the data is greater than this cutoff. We do this using a condition analogous to equation 18, and we continue the above process sequentially for $\mathcal{H}=1,...,\mathcal{H}_{\max}$ until we fail to conclude that there are $\mathcal{H}$ or more production technologies. In particular, we conclude that there are H production technologies according to the following criteria:

$$\text{Conclude } H \text{ production technologies} \iff R_{H|H-1} < R^{\hat{\text{syn}}}_{H|H-1} \ \& \ R_{H+1|H} \geq R^{\hat{\text{syn}}}_{H+1|H}. \tag{19}$$

For example, we would conclude that there are three production technologies if we find evidence both in favor of there being three or more production technologies and against there being four or more production technologies.

We verify the validity of our methodology; that is, we recover the true number of production technologies and assign each firm to the correct production technology, using simulations. We generate samples that have between one and three production technologies, and we find that our methodology correctly recovers the true number of production technologies in 100 percent of our simulations and assigns all firms to the correct production technology for a range of plausible parameter values. We explore the robustness of our methodology when equation 14 is not satisfied and find that our methodology still works well in recovering the number of production technologies, even though a small number of individual firms may be misclassified. The full details of how we evaluate our methodology through simulations are available in the appendix.

3.4 Recovering Unobservable Parameters

After identifying the total number of production technologies following Section 3.3, we are able to use our clustering algorithm from Section 3.2 to determine the production technology used by each individual firm. In terms of our model parameters, this gives us H and the mapping $\hbar(m)$ that tells us which of the H technologies is used by firm m. The next step is to use these objects to identify the remaining unobservable parameters in our framework: factor intensities, markups, and distortions.

The crux of our strategy for recovering the unobserved parameters relies on using equation 11 and exploiting variation in cost shares. We cannot disentangle factor intensities from the average distortion level without making additional assumptions. Therefore, we follow the misallocation literature by assuming that distortions are centered on the undistorted allocation. Specifically, we assume that the mean of the distribution of inverted idiosyncratic distortions described in Section 3.2 is equal to 1, that is,

$$\frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \left(\tau_m^i\right)^{-1}}{M_h} = 1, \forall h \in \{1, ..., H\}, \forall i \in \{1, ..., I\}, \tag{20}$$

where $M_h \equiv \sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h}$ is the number of firms that use technology h. Similarly, we assume that the average of the inverted technology-specific distortions is equal to 1,

$$\frac{1}{H} \sum_{h=1}^{H} (\tau_h)^{-1} = 1. \tag{21}$$

We view these assumptions on the mean value of distortions as innocuous. This is because distortions lead to misallocation by altering the allocation of factor inputs across firms. Therefore, multiplying all distortions by a constant positive number does not ultimately change allocations. These assumptions do not rule out misallocation across production technologies. First, we are able to capture differences in the average level of distortions across production technologies through our technology-specific distortion, τ_h. Second, distortions may be correlated with firm productivity, and firm productivity may be correlated with a firm's production technology, which can lead to misallocation across production technologies.

We discuss our methodology for recovering each of the remaining parameters in the following subsections. The basic idea is to use equation 11, and if a parameter does not vary across inputs, then averaging cost shares across inputs will help us to pin down the parameter. Likewise, if a parameter does not vary across firms, we will want to average cost shares over firms.

3.4.1 Markups

Markups are determined by the elasticity of substitution in equation 9 and all firms in an industry charge the same constant markup over marginal cost. To derive an expression for markups, we sum equation 11 for a given input i across all firms using technology h:

$$\frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \tilde{\theta}_m^i}{M_h} = \rho \left(\tau_h\right)^{-1} \theta_h^i. \tag{22}$$

In the above equation, we removed idiosyncratic distortions by plugging in our assumption from equation 20. We next exploit the constant returns to scale nature of the production function by summing equation 22 across all inputs for firms that use technology h to get

$$\sum_{i=1}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \tilde{\theta}_m^i}{M_h} = \rho \left(\tau_h\right)^{-1} \sum_{i=1}^{I} \theta_h^i. \tag{23}$$

From constant returns to scale we have $\sum_i \theta_h^i = 1$, which we can plug into the above equation. We can rearrange and sum equation 23 over all technologies to get the following:

$$\rho \sum_{h=1}^{H} \left(\tau_h\right)^{-1} = \sum_{h=1}^{H} \sum_{i=1}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \tilde{\theta}_m^i}{M_h}. \tag{24}$$

The above reveals our need for an assumption on the average technology-specific distortion in order to compute markups. In particular, we use the condition from equation 21 to arrive at

$$\rho = \frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \tilde{\theta}_m^i}{M_h}, \tag{25}$$

which allows us to determine the markup for firms in an industry. This is because $\tilde{\theta}_m^i$ is directly observable, and we know $\hbar(m)$ and M_h from our clustering methodology in Sections 3.2 and 3.3. For an example of how to interpret the above equation, suppose that observed expenditures on inputs are a small fraction of revenue on average within an industry. Equation 25 implies that ρ will be low, which is consistent with high markups in the industry.

3.4.2 Technology-Specific Distortions

To recover the technology-specific distortion, τ_h, we exploit that it does not vary across factor inputs or across firms that share a production technology. Therefore, we can rearrange equation 23, where we have averaged across firms within a given production technology and across inputs, to derive the following relationship:

$$\tau_h = \rho \left(\sum_{i=1}^{I} \frac{\sum_{m=1}^{M} \mathbb{I}_{\hbar(m)=h} \tilde{\theta}_m^i}{M_h} \right)^{-1}. \tag{26}$$

From this equation we can compute the technology-specific distortions by using our estimate of ρ from equation 25. The intuition of the expression is that if observed expenditures are a smaller fraction of revenues for firms in one production technology, then τ_h will be high which means that firms using that production technology face higher input costs.

3.4.3 Factor Intensities

Our methodology in Sections 3.2 and 3.3 allows us to identify the production technology used by each firm; however, we have not yet determined the factor intensities associated with each production technology. To determine the factor intensities for each production technology, we average cost shares across firms within each production technology and then adjust for the parameters that affect the mean cost share. We rearrange equation 22 to get the following expression:

$$\theta_h^i = \frac{\tau_h}{\rho} \frac{\sum_{m=1}^{M} \mathbb{I}_{\hat{h}(m)=h} \tilde{\theta}_m^i}{M_h}. \tag{27}$$

Given estimates of τ_h and ρ, we can then calculate the factor intensity θ_i^h for input i and production technology h. We require an estimate of τ_h because a higher technology-specific distortion will deflate the mean cost share within a production technology. We require an estimate of ρ to adjust for markups because our strategy relies on using cost shares (factor expenditures divided by total revenue) rather than simply factor expenditures divided by total expenditures. Higher markups therefore deflate the ratio of expenditures to total revenues on any given factor.

3.4.4 Idiosyncratic Distortions

The final unobserved variables we have to recover are the idiosyncratic firm-level distortions, τ_m^i, for each firm m and input i. We find these firm-level distortions by rearranging equation 11 as follows:

$$\tau_m^i = \frac{\rho \theta_h^i}{\tau_h \tilde{\theta}_m^i}. \tag{28}$$

The above expression closely resembles the strategy followed by previous studies in the misallocation literature. Specifically, τ_m^i is the wedge necessary to rationalize the difference between observed cost shares and our recovered factor intensities. In the absence of distortions, firm optimization would require the fraction of expenditures on each input to be equal to the factor intensity, meaning that cost shares would satisfy $\tilde{\theta}_m^i = \rho \theta_h^i$. Deviations from this relationship are the standard way of recovering distortions. These expressions highlight that variation in cost shares among firms can only be attributed to distortions if they share a common production technology. In contrast, if we only allow for a single production technology, all variation in cost shares would be attributed to distortions.

4 Application to Chilean Plant-Level Data

To examine the pervasiveness of multiple production technologies in practice, we now apply our methodology from Section 3 to Chilean plant-level data.

4.1 Data Description and Preparation

We use the 2005 vintage of the Encuesta Nacional de Industria Anual (ENIA) dataset collected by the Instituto Nacional de Estadisticas (INE), which covers all active manufacturing plants in Chile with more than 10 employees. A benefit of the ENIA plant-level data is that the data have been thoroughly examined by previous studies, including studies on misallocation and on estimating production functions.[7] This information is useful since the reliability of the Chilean data is well understood and established from these previous studies.

The ENIA dataset classifies each plant according to its four-digit ISIC Revision 3 industry code, which is what we use as our definition for an industry. Along with each plant's industry, we require information on revenues as well as expenditures on capital, unskilled labor, skilled labor, and intermediate inputs. Previous studies on misallocation have primarily used only capital and labor; however, our inclusion of intermediate inputs and unskilled labor separately from skilled labor is useful for distinguishing when firms employ different production technologies by giving us more margins along which to apply our clustering algorithm. Future studies may benefit from breaking down intermediate input usage even further — for example by separating electricity usage, fuel usage, and water usage — to help identify the various production technologies within an industry. We refrain from including these inputs separately, however, because of the prevalence of missing observations for these inputs in many industries.[8]

For revenues we use the nominal gross output reported by plants. For expenditures on skilled and unskilled labor, we use nominal labor remuneration for each type of labor. For intermediate inputs, we use the nominal value reported by plants. Unlike expenditures on intermediate inputs and labor, we do not have a direct measure of the user cost of capital. We utilize a strategy similar to that in Young (1995) to recover the user cost of capital by using the following no-arbitrage condition for each type of capital that we consider:

$$R_{tj} = 1 + r_t - (1 - \delta_j)\,\frac{P_t}{P_{t+1}}\,\frac{P^K_{t+1}}{P^K_t}, \tag{29}$$

where R_{tj} is the user cost of capital type j, δ_j is the depreciation rate of capital type j, P_t is

[7] Papers that have used Chilean plant-level data include Alvarez and López (2005), Asturias, Hur, Kehoe, and Ruhl (2017), Bergoeing and Repetto (2006), Levinsohn (1999), Levinsohn and Petrin (2003), Oberfield (2013), and Pavcnik (2002), Petrin and Levinsohn (2012).

[8] Care should be taken to avoid including inputs arbitrarily. For instance, if we include inputs that would not be expected to display constant expenditure shares, this would violate our assumption of Cobb-Douglas production functions and may lead us to incorrectly estimate the number of production technologies.

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/P_{t+1}, and the investment deflator to determine P_{t+1}^K/P_t^K, all of which we obtain from the World Development Indicators database. We assume depreciation rates of 20 percent for vehicles, 10 percent for machinery, and 5 percent for buildings, which is consistent with that in Levinsohn and Petrin (2003). We multiply the user cost of capital for each type of capital, found using equation 29, by the nominal value of that capital. To determine the nominal value of each type of capital, we use the plant's reported book value of capital for vehicles, machinery, and buildings.

4.2 Clustering Results

We apply our methodology described in Section 3 to the plant-level data described above. For the analysis, we apply our methodology separately for each industry. Our methodology requires a sufficient number of plants for each production technology. Therefore, we exclude industries with fewer than 50 plants, and we cap the potential number of production technologies per industry by setting $\mathcal{H}_{\max} = 4$. After we exclude industries with insufficient observations, we are left with 23 industries containing a total of 2,754 plants. For each industry, our methodology yields the number of production technologies, which production technology is used by each plant, the industry markup over marginal cost, factor intensities for each production technology, and distortions for each plant. In order to place our results in context, we also conduct an analysis where we back out each object under the assumption that there is a single production technology in each industry. We do this by repeating the steps from Section 3.4 under the assumption that $H = 1$ and $\hbar(m) = 1$ for each plant m.

Table 1 reports the industries included in our analysis, the number of plants operating in each industry, and the number of production technologies identified by our methodology. We find multiple production technologies in a majority of the industries in our sample: 14 industries have multiple technologies present compared to 9 industries for which we cannot identify more than a single production technology. There is no relationship between the number of plants in an industry and the number of production technologies we recover.

As an example of an industry with multiple production technologies, consider the manufacturing of non-ferrous and precious metals industry (ISIC 2720). In Chile this industry is primarily composed of plants involved in copper processing (we will henceforth refer to ISIC 2720 as the copper manufacturing industry). It is the largest industry by gross output in our dataset and exports of copper products accounted for over 50 percent of Chile's exports in 2005 according to data from the United Nation's Comtrade database.

Our methodology finds that two production technologies are operating in the copper manufacturing industry. We find that 19 plants operate what we refer to as technology 1, while 43 plants operate technology 2. This finding reflects widespread use of both technologies. Table 2 describes the recovered factor intensities when we allow for multiple production technologies and compares them to those obtained under the standard assumption of a single technology. We find that when we only allow for a single production technology, the implied factor intensities are effectively a weighted average of the factor intensities of the two production technologies. For example, we find that technology 1 has a factor intensity of 0.87 for intermediate inputs and technology 2 has a factor intensity of 0.78. If we only allow for one production technology, then the inferred factor intensity is 0.80.

Figure 1

Kernel Density Plot of Gross Output
Copper Manufacturing Industry

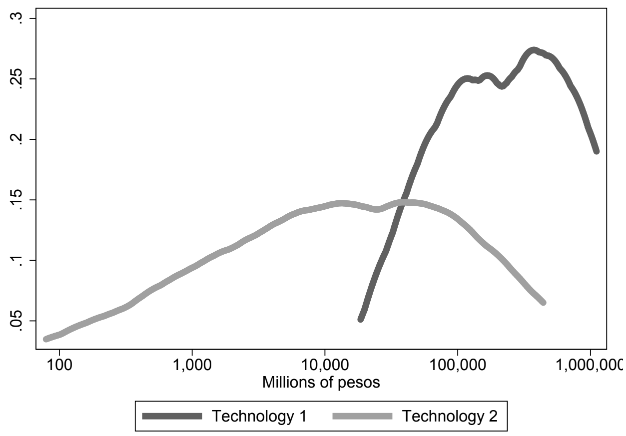


Figure 1 shows the kernel density plot of the gross output of plants in the copper manufacturing industry by the recovered technology of plants.

It is interesting to explore how the characteristics of firms vary across production technology usage. We find that there are significant differences in plant size between these two technologies. Figure 1 shows the kernel density plots by gross output by technology. We see that technology 1 has significantly larger plants. In fact, the median plant that uses technology 1 has a gross output that is 14.9 times that of the median plant using technology 2.

We now investigate the relationship between plant size and production technology usage across all the plants in our dataset. To do so, we regress the log of gross output of plants on the log of the factor intensity of an input based on the plant's technology. We include industry fixed effects, in order to compare the factor intensities of large and small plants within the same industry. We estimate this regression separately for each of the four inputs and report the results of these regressions in Table 3. We find that larger plants tend to have technologies that use intermediate inputs more intensively. These larger plants also tend to use skilled and unskilled labor less intensively, which is consistent with our findings for the copper manufacturing industry. Conversely, we do not find a significant difference in capital intensity across large and small plants.

It is useful to consider the implications of these findings in terms of the relative intensities of capital and labor. As mentioned before, Table 3 shows that larger plants have technologies that use labor less intensively and use capital with the same intensity. This finding implies that the larger plants' capital expenditure is higher relative to their labor expenditure. Therefore, when we view these results through the lens of a value-added production function, larger plants tend to use capital-intensive technologies whereas smaller plants tend to have labor-intensive technologies. In the case of the copper manufacturing industry, for example, if we rewrite the production technologies found in Table 2 in value-added terms, the coefficient of capital is 0.71 for technology 1 and 0.32 for technology 2.

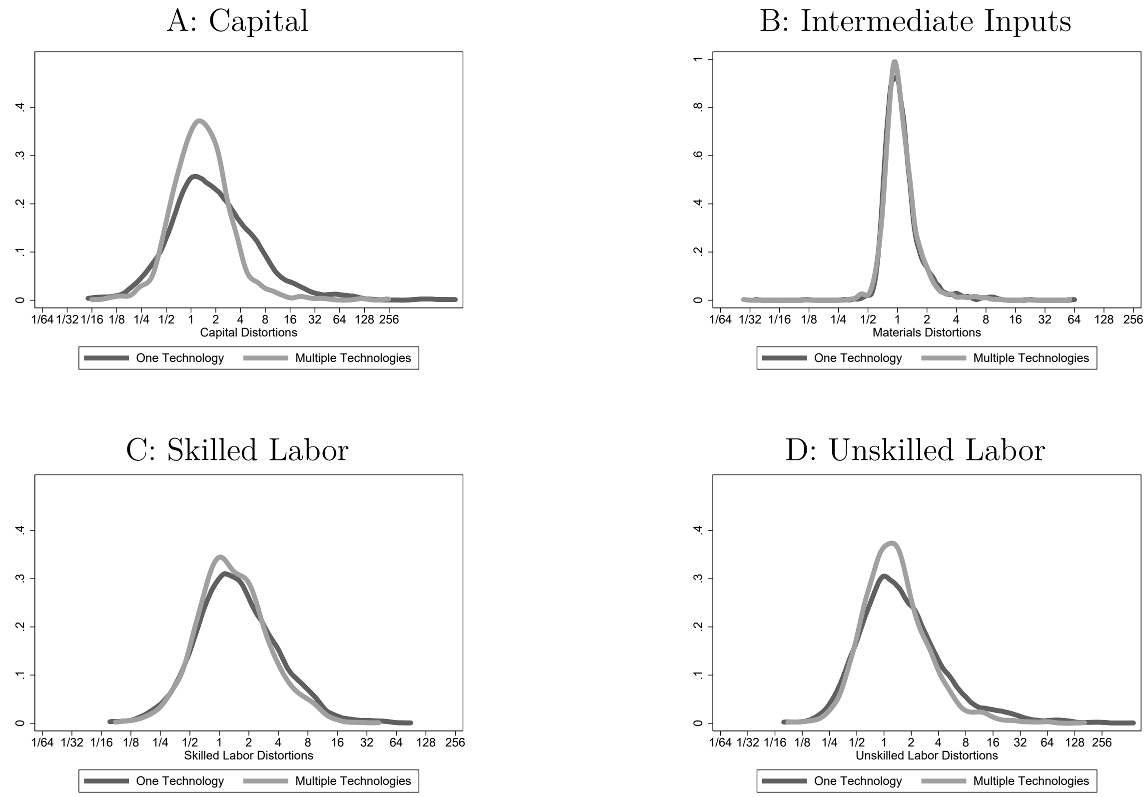
4.3 Dispersion in Marginal Revenue Products

We now examine how the dispersion in marginal revenue product for each input changes after we account for the presence of multiple production technologies. In our framework, dispersion in marginal revenue products necessarily implies that there is misallocation, since it would be efficient to reallocate inputs from plants with low to high marginal revenue product until marginal revenue products are equalized across firms. As mentioned in Section 2.1, the marginal revenue product of input i for firm m is $\tau_h \tau_m^i p^i$. Following the misallocation literature, we focus on the dispersion of logged marginal revenue products. We can ignore input prices in our analysis without affecting our chosen measure of dispersion, the standard deviation of logged marginal revenue products, because we assume p^i is constant for all firms within an industry. In Section 4.4 we impose additional model structure in order to allow us to compute input prices, but for now it is useful that we can explore dispersion in marginal revenue products without imposing additional structure.

Table 4 lists the percentage reductions in the standard deviations of logged marginal revenue products for each of the four inputs after allowing for multiple production technologies. We report the reduction for each of the industries we identified as having more than one production technology in Table 1. We also report the aggregate reduction in the dispersion of marginal revenue products for each input, which is the reduction in the standard deviation when we jointly consider all plants in the economy across all industries regardless of the number of production technologies. The aggregate measures require assuming that input prices are constant across industries. We find significant variation in the magnitude of the reduction across industries and inputs. Overall we find the largest reductions for capital, with an aggregate decline of 15.0 percent. We also see reductions in the dispersion of the marginal revenue product for unskilled labor of 12.1 percent and for skilled labor of 6.8 percent. There are minimal reductions in dispersion for intermediate input usage, in which we saw an aggregate decline of only 1.3 percent and small increases in some industries.

Figure 2

Kernel Density Plots of Marginal Revenue Product of Inputs One vs. Multiple Technologies



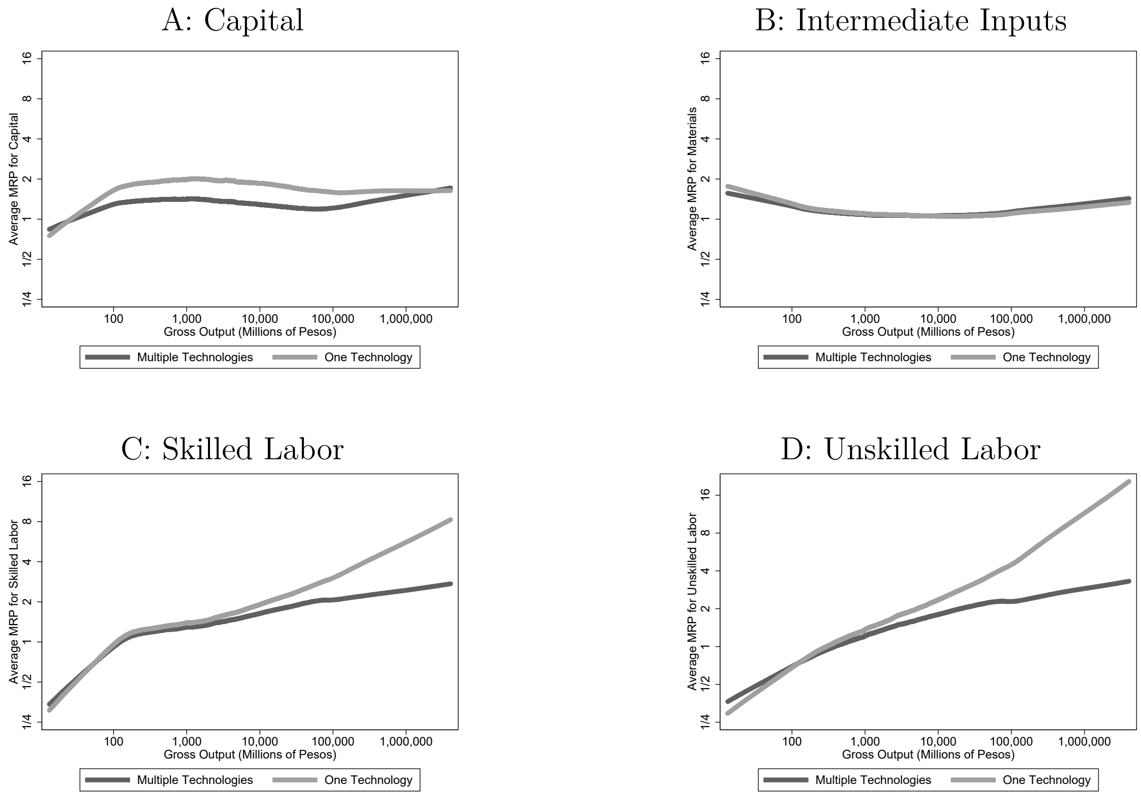
Panel A of Figure 2 shows the kernel density plot of the marginal revenue product of capital under one technology and multiple technologies; Panel B shows the same plot for intermediate inputs; Panel C shows the same plot for skilled labor; Panel D shows the same plot for unskilled labor. We use a Epanechnikov kernel function for the kernel density plot.

Figure 2 displays the kernel density plots of the marginal revenue products for each input for all plants in our sample. We display the kernel density functions under the assumption of a single production technology and after allowing for multiple production technologies. The kernel density plots serve to flexibly characterize the distributions of the marginal revenue products by non-parametrically estimating the probability density functions for each of the four inputs. From these figures we can observe that the dispersion in the marginal revenue products of capital exhibits the largest decline, which is consistent with the results of Table 4. Conversely, there is relatively little dispersion in the marginal revenue product of intermediate inputs under both specifications. This finding is consistent with the findings of Petrin and Sivadasan (2013), who find low levels of misallocation in intermediate inputs among Chilean plants using ENIA plant-level data covering the period 1982-1994.

Restuccia and Rogerson (2008) point out that the correlations between idiosyncratic distortions and the size of firms can play an important role in determining the impact these distortions have on overall misallocation. In particular, if the marginal revenue product of an input is increasing with firm size, this indicates large potential gains from eliminating distortions. Eliminating distortions would then lead to the reallocation of the input from smaller firms to large firms, which tend to be more productive. In our framework, this same insight applies, while at the same time we are able to recover systematic differences in the production technologies used by firms of different sizes. Thus, it is informative to study how the relationship between the marginal revenue product of inputs and size changes if we consider multiple technologies.

Figure 3

Marginal Revenue Product of Inputs by Gross Output
One vs. Multiple Technologies



Panel A of Figure 3 shows the lowess fitted curve for the marginal revenue product of capital by gross output under one technology and multiple technologies; Panel B shows the same plot for intermediate inputs; Panel C shows the same plot for skilled labor; Panel D shows the same plot for unskilled labor. We use least square smoothing and a bandwidth of 0.8 for the lowess locally weighted regression.

In Figure 3 we plot the relationship between the marginal revenue products of each input and the gross output of plants when we consider one and multiple technologies. These figures are constructed by fitting the data with a lowess locally weighted regression, which allows us to non-parametrically and flexibly characterize the relationship between marginal revenue products and size. To get a sense of the scale of plants, 100 million pesos is equivalent to approximately $180,000 US dollars using the average 2005 exchange rate provided by the World Development Indicators database.

We find a strong positive relationship between size and the marginal revenue product of both skilled and unskilled labor. This finding suggests that large plants use too little skilled and unskilled labor, and therefore it would be efficient to reallocate both types of labor from small to large plants. The correlation between size and marginal revenue product is positive both with multiple production technologies and with a single production technology; however, forcing a single production technology significantly overstates the magnitude of the correlation. For example, after accounting for multiple production technologies, the predicted marginal revenue product of skilled labor for plants in the 95th percentile of size is 2.44 times that of the plant that is in the 5th percentile of size. Forcing a single production technology overstates this ratio by over 30 percent, yielding a ratio of 3.20. For unskilled labor, the correlation and the overstatement are both significantly larger. The ratio is 4.50 with multiple production technologies and 8.24 for a single production technology, meaning the ratio is overstated by over 80 percent if we only use a single production technology. Our results highlight that large firms face a higher marginal cost for hiring unskilled labor, but also that large firms are more likely to employ production technologies that require less unskilled labor. Ignoring that large firms use systematically different production technologies than small firms leads to overstating the correlation between size and marginal revenue products.

For capital and intermediate inputs, we find much weaker relationships between size and marginal revenue products, with ratios of 1.27 and 1.23, respectively, for the marginal revenue products of capital and intermediate inputs for firms at the 95th percentile of size versus the 5th percentile. We similarly observe minimal change in these ratios after accounting for multiple production technologies. For capital, this is because we do not observe large firms using systematically more or less capital-intensive technologies than small firms. For intermediate inputs, we do see that larger firms use production technologies with higher intermediate input intensities; however, the overall effect on the marginal revenue products for intermediate inputs is small because the factor intensities for intermediate input usage tend to be large. For example, in Table 2 in the single production technology case, the factor intensity for unskilled labor is 0.05, whereas the factor intensity for intermediate inputs is 0.80. This means the difference in estimated factor intensities after allowing for multiple production technologies must be roughly 16 (0.80/0.05) times larger for intermediate inputs than for unskilled labor to yield the same overstatement in marginal revenue products. Conversely, a 10 percent distortion to the price of intermediate inputs would have a much larger impact than the same distortion applied to unskilled labor; so this information does not imply that misallocation in intermediate inputs is unimportant. In our framework, misallocation in intermediate inputs ends up being a non-negligible part of overall misallocation despite low dispersion in marginal revenue products arising from the importance of intermediate inputs in the production functions we estimate for firms.

4.4 Gains from Eliminating all Distortions

We now want to understand the impact of incorporating multiple technologies on the estimated gains in manufacturing output and TFP from eliminating all distortions in the economy.[9] Our results from the previous section highlight two main reasons why accounting for multiple production technologies would be expected to reduce the predicted gains from misallocation. The first reason is that accounting for multiple production technologies reduces the estimated dispersion in marginal revenue products of inputs across firms, particularly for capital, suggesting smaller distortions and therefore smaller gains from eliminating them. The second reason is that accounting for different technology usage across large and small firms can explain a large part of the relationship between marginal revenue products and the size of firms for skilled and unskilled labor. By reducing the magnitude of this association, we can be expected to reduce the predicted gains from eliminating distortions.

We want to allow for reallocation across industries; therefore, we impose additional structure on the model presented in Section 2. In particular, we consider a multi-industry extension where the output of each industry (let the subscript s denote the industry for each variable from Section 2) is used to create a single final consumption good, Y, by a perfectly competitive bundler with the following Cobb-Douglas production function:

$$Y = \prod_{s=1}^{S} (Y_s)^{\alpha_s}, \tag{30}$$

where Y_s is the output from industry s, which we get from equation 2, α_s is industry s's share of total expenditures, and S is the total number of industries in the economy.

We assume that factors are fully mobile across industries so that the income received for supplying one unit of each input will be the same regardless of which industry or firm uses that input. The budget constraint of the representative consumer implies that total expenditures will be equal to income from factor inputs, $p^i X^i$ for input i, profits from firms, Π, and transfers from firms due to distortions, T. Total expenditures across all industries in the economy are therefore given by

$$PY = \sum_{i=1}^{I} p^i X^i + \Pi + T, \tag{31}$$

where X^i denotes the exogenously determined and inelastically supplied total amount of factor i in the economy and P is the aggregate price index:

$$P = \prod_{s=1}^{S} \left(\frac{P_s}{\alpha_s}\right)^{\alpha_s}, \tag{32}$$

[9] We follow the typical definition of misallocation in the literature and define it as losses that result from the presence of distortions. In our framework, eliminating markup heterogeneity across industries would generate an increase in aggregate output; however, we do not classify this as misallocation or remove markup heterogeneity in our counterfactuals. The papers by Peters (2013) and Asturias, García-Santana, and Ramos (2016) are recent examples of studies that examine the losses in output that result from markup heterogeneity.

where P_s is defined in equation 4. Profits are the result of markups over marginal cost and are given by

$$\Pi = \sum_{s=1}^{S} \left(\frac{1-\rho_s}{\rho_s} \right) \left(\sum_{m=1}^{M_s} p_{m,s} y_{m,s} \right). \tag{33}$$

Note that markups are allowed to vary across industries. Transfers are defined as

$$T = \sum_{s=1}^{S} \sum_{m=1}^{M_s} \sum_{i=1}^{I} \left(\tau_{h,s} \tau_{m,s}^{i} - 1 \right) p^i x_{m,s}^{i}. \tag{34}$$

The clearing condition for each input i is given by

$$X^i = \sum_{s=1}^{S} \sum_{m=1}^{M_s} x_{m,s}^{i}, \tag{35}$$

where the quantity of input demanded by each firm is given by

$$x_{m,s}^{i} = \theta_{\hbar(m),s}^{i} \frac{\frac{1}{\rho_s} p_{m,s} y_{m,s}}{\tau_{\hbar(m),s} \tau_{m}^{i} p^i}. \tag{36}$$

The previous equation shows that firms that face a high distortion for a given input will demand less of that input in equilibrium. In the absence of distortions, we would expect constant expenditure shares across firms that share a common production technology.

In order to calculate the gains from eliminating distortions, we need to calibrate the productivity of each plant. We are interested in estimating the percentage change in output given the removal of distortions, which means that we only need to know the relative productivities of firms within each industry. To calculate relative productivities, we first normalize the productivity of one plant, $\tilde{m}$, in each industry to 1. Let h represent the technology of firm m and $\tilde{h}$ represent the technology of firm $\tilde{m}$, both in industry s. We then combine equations 3 and 6 to find the productivity of plant m relative to that of $\tilde{m}$. We have that the productivity of plant m relative to plant $\tilde{m}$ is

$$\frac{z_{m,s}}{z_{\tilde{m},s}} = \left(\frac{p_{m,s} y_{m,s}}{p_{\tilde{m},s} y_{\tilde{m},s}} \right)^{\frac{1-\rho_s}{\rho_s}} \frac{\prod_{i=1}^{I} \left(\frac{\tau_{m,s}^{i} p^i}{\theta_{\hbar(m)}^{i}} \right)^{\theta_{\hbar(m)}^{i}} \left(\tau_{\hbar(m),s} \right)^{-1}}{\prod_{i=1}^{I} \left(\frac{\tau_{\tilde{m},s}^{i} p^i}{\theta_{\hbar(\tilde{m})}^{i}} \right)^{\theta_{\hbar(\tilde{m})}^{i}} \left(\tau_{\hbar(\tilde{m}),s} \right)^{-1}}, \tag{37}$$

where we normalize $z_{\tilde{m},s} = 1$ to find $z_{m,s}$. Note that if firms share a production technology, their relative productivity is a function of their relative revenues and a weighted geometric mean of their relative distortions, where the weights of the geometric means are the factor intensities.

When we calibrate our model, revenues and input expenditures come from the data; however, our data do not break down values into prices and quantities. We are only interested in changes in output; therefore, we can normalize the total supply of each factor equal to one, that is, set $X^i = 1$, for all i. This normalization does not alter our results because it gets absorbed into prices and therefore into how we calibrate our productivities. If we instead were able to identify the productivity of each firm independently of calibrating it using equation 37, then changing the total supply of each factor could lead to different results in terms of the gains from reducing misallocation. We can show, for example, that the relative amount of inputs across firms does not change with the total endowment of that input in our calibration. Equation 36 shows that the demand for an input is a function of the revenues of a firm. We take firm revenues directly from the data in our calibration; therefore, the relative amount of inputs across firms remains the same even if we change the total endowment. Observed factor expenditures, $p^i x^i_{m,s}$, are likewise taken from the data, which highlights that an increase in the endowment of an input must be absorbed into an equal and inverse decrease in the price of that input.

After normalizing the total supply of each factor, we can calculate the equilibrium numerically by solving equations 30–37 for input prices, productivities, and output of each firm and therefore output for each industry, Y_s, and aggregate output, Y. We repeat the same exercise after setting $\tau_{\hbar(m),s} = 1$ and $\tau^i_{m,s} = 1$ for all firms, industries, and inputs. This allows us to predict aggregate output in the absence of distortions $\hat{Y}$. We calculate the percentage gain in aggregate output from eliminating misallocation as

$$\Delta Y \equiv 100 * \frac{\hat{Y} - Y}{Y}. \tag{38}$$

We similarly calculate the percentage gain in output, ΔY_s, for each industry, s, by substituting industry-level output, Y_s and $\hat{Y}_s$, for aggregate output in the above equation.

We calculate the gains from eliminating misallocation under two scenarios. First, we calculate the gains in aggregate output using our methodology to recover the number of production technologies in each industry. Second, we calculate the gains under the standard assumption of a single production technology in each industry, which we do by setting $H = 1$ and then following the rest of our methodology above and in Section 3.4. Table 5 reports the percentage changes in industry-level output and aggregate output from eliminating misallocation. We find that if we assume each industry has a single technology then the predicted gains in aggregate output are 38.7 percent. After we allow for multiple production technologies, then the predicted gains from eliminating misallocation drop to 26.8 percent — a reduction of nearly 1/3. Thus, we find significant declines in the predicted output gains from the elimination of misallocation after we allow for multiple production technologies. Whereas most industries experience a growth in output, we also find a decline in output after the removal of distortions for some industries. This result highlights the importance of the reallocation of inputs across sectors.

To understand where the largest gains in productivity arise at the industry level, we conduct a similar exercise for changes in TFP. Table 6 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 suppressed for the technologies) is

$$Y_{h,s} \equiv \left(\sum_{m=1}^{M_s} \mathbb{I}_{\hbar(m)=h} \left(y_{m,s}\right)^{\rho_s} \right)^{1/\rho_s}. \tag{39}$$

Thus, the TFP for technology h is

$$A_{h,s} = \frac{Y_{h,s}}{\prod_{i=1}^{I} \left(X_{h,s}^{i}\right)^{\theta_h^i}}, \tag{40}$$

where X_h^i is the total amount of input i — capital, skilled labor, unskilled labor, and intermediate inputs — used by firms utilizing technology h, which is characterized by

$$X_{h,s}^{i} \equiv \sum_{m=1}^{M_s} \mathbb{I}_{\hbar(m)=h} x_{m,s}. \tag{41}$$

We find the percentage change in TFP for each technology after the removal of distortions, $A'_{h,s}$. To calculate changes in industry-level TFP, we weight these percentage changes with the share of gross output accounted for by firms that use that technology in the calibrated economy:

$$\Delta A_s \equiv \sum_h \left(\frac{A'_{h,s}}{A_{h,s}} P_{h,s} Y_{h,s} \right). \tag{42}$$

Table 6 reports the change in TFP for each industry following the elimination of misallocation. Further, it reports the change in aggregate TFP. We compute changes in aggregate TFP as follows:

$$\Delta A \equiv \sum_s \left(\alpha_s \Delta A_s \right),$$

where the change in aggregate TFP, ΔA, is defined as the weighted average of changes in industry TFPs, ΔA_s, where the weight for each industry is its Cobb-Douglas expenditure share, α_s. We find that the aggregate gains in TFP are 24.3 percent under multiple technologies and 30.9 percent if we only allow for one technology. As before, we find a significant reduction — nearly 1/4 — in the predicted gain from eliminating misallocation in the economy. This reduction takes into account the TFP of industries with only a single production technology. If we focus on changes in industry-level TFP only in industries with multiple technologies our results are similar, with a mean reduction of 27.2 percent.

5 Conclusion

In this paper, we proposed a methodology for using cluster analysis to identify heterogeneous technology usage within industries, where technologies are represented by Cobb-Douglas production functions with differing factor intensities. We motivate the use of cluster analysis by connecting it to an economic model featuring unobserved distortions that cause variation in observed cost shares across firms. Importantly, our methodology requires only data on revenues and factor input expenditures and works by identifying gaps or areas of low density in the multi-dimensional distributions of cost shares. We provide a sufficient condition showing that when technologies are sufficiently distinct relative to the variation in distortions, clustering on cost shares will correctly group firms by their production technology. Additionally, we provide evidence from simulations to show that our methodology works well in recovering unobserved production technologies and to explore its robustness to alternative specifications.

We apply our methodology to data on Chilean manufacturing plants, which we use to calculate the gains in output and TFP from eliminating misallocation through the removal of unobserved distortions. Whereas the standard assumption in the misallocation literature is that all firms within an industry employ a common production technology, we find strong evidence of multiple production technologies in the majority of industries. We show that after accounting for firms employing different production technologies, the predicted gains from eliminating misallocation decline by nearly a third. We further find significant reductions in the variance of marginal revenue products of factor inputs, particularly for capital and unskilled labor, compared to standard estimates that assume a single production technology for all firms. In addition, we find that adjusting for differences in production technologies across firms reduces the degree to which marginal revenue products are higher for large firms than for small firms. Our results show that a non-trivial portion of measured misallocation can be attributed to measurement error arising from assuming that firms with different production technologies employ a common production technology. We expect our methodology will be helpful for researchers searching for specific policies or sources of measurement error that might explain observed variation in marginal revenue products of inputs across firms.

Our methodology for determining the number of production technologies is robust to moderate amounts of measurement error that operate through increasing dispersion in cost shares. This measurement error will still be incorporated into our recovered distortion. This is a common shortcoming with misallocation studies, although it should be less of an issue for our paper because our focus is not on the level of misallocation itself, but what fraction of measured misallocation (which includes measurement error) can be accounted for by incorporating multiple production technologies.

In this paper we utilized the k-means++ clustering algorithm, which assigns each firm to a single production technology. This choice makes it straightforward to conduct counterfactuals when using our methodology to compute the gains from eliminating misallocation. For other applications, researchers may be interested in exploring alternative clustering and machine learning algorithms. For example, mixed-membership models are used by search engines to categorize objects such as news articles which may fall into multiple categories and could be useful for analyzing multi-product plants that employ multiple technologies simultaneously. One of the primary contributions of this paper is the use of an economic model to justify which variables the clustering algorithm should be applied to, in what cases the methodology will be successful, and how to use the results of the algorithm to recover specific unobservable parameters in the original model. Therefore, although we expect these algorithms to be useful for analyzing the increasingly disaggregated plant-level data available to researchers, we also want to emphasize the importance of connecting the algorithms employed to specific economic models, as we do in this paper.

Another area for future research is understanding the ways in which firms select their production technologies. In this paper, we considered production technologies to be exogenously given and non-responsive to changes in distortions. In practice, firms may alter their technology usage in response to changes in their environment, such as changes in prices or distortions, as well as adjustment costs in selecting a different production technology. This implies that along with the misallocation that exists conditional on each firm's production technology — which we study in this paper — there could potentially be misallocation in terms of the production technologies firms are utilizing. We expect our methodology will be useful for exploring technology misallocation by allowing researchers to identify the production technologies at use within an industry and how technology usage has changed over time, particularly when combined with data on firm-to-firm linkages and intermediate input usage. In the longer run, we think it is important to continue to develop methodologies that help researchers investigate whether specific policies can encourage development by facilitating changes in production technologies.

Table 1

Number of Production Technologies by Industry

ISIC	Industry Description	Firms	Prod. Technologies
1511	Processing and preserving of meat products	113	2
1512	Processing and preserving of fish products	152	3
1513	Processing and preserving of fruits and vegetables	80	1
1520	Manufacture of dairy products	56	1
1531	Manufacture of grain mill products	88	1
1541	Manufacture of bakery products	499	1
1552	Manufacture of wines	110	4
1711	Preparation and weaving of textiles	53	3
1810	Manufacture of wearing apparel	188	1
1920	Manufacture of footwear	60	1
2010	Sawmilling and planing of wood	196	2
2102	Manufacture of corrugated paper and paperboard	63	3
2211	Publishing of books, brochures, and other publications	52	1
2221	Printing	64	4
2411	Manufacture of basic chemicals	66	3
2424	Manufacture of soap, cleaning preparations, and perfumes	58	2
2429	Manufacture of other chemical products n.e.c.	61	2
2520	Manufacture of plastics products	242	2
2695	Manufacture of articles of concrete, cement and plaster	103	2
2720	Manufacture of basic precious and non-ferrous metals	62	2
2811	Manufacture of structural metal products	120	3
2899	Manufacture of other fabricated metal products n.e.c.	125	1
3610	Manufacture of furniture	143	1

Table 1 shows the number of technologies determined by the clustering methodology at the industry-level. *ISIC* is the four-digit ISIC Rev. 3 code for each industry we evaluate with our methodology (>50 firms). *Industry Description* lists a shortened description of the economic activities that make up each industry. *Firms* lists the number of firms (plants) in each industry. *Prod. Technologies* shows the number of production technologies operating in each industry as recovered by our methology.

Table 2

Recovered Production Technologies

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

Case	Technology	Count	θ_{sl}	θ_{ul}	θ_k	θ_x
Single Technology		62	0.07	0.05	0.08	0.80
Multiple Technologies	1	19	0.03	0.01	0.10	0.87
Multiple Technologies	2	43	0.08	0.07	0.07	0.78

Table 2 shows the recovered factor intensities for the manufacturing of non-ferrous and precious metals (ISIC 2720). *Case* indicates whether we report the case in which we impose a single technology or allow for multiple technologies. *Technology* assigns each technology an identifying number, which is useful if there is more than one technology in the industry. These identifiers are used in Figure 1. *Count* lists the number of plants that use each technology. Columns 4-7 report the factor intensities for skilled labor, unskilled labor, capital, and intermediate inputs, respectively.

Table 3

Factor Intensities vs. Gross Output

	Logged Factor Intensity			
	Capital	Skilled Labor	Unskilled Labor	Intermediate Inputs
Logged Gross Output	-0.0106	-0.0382***	-0.0846***	0.0119***
	(0.0130)	(0.00590)	(0.00892)	(0.00206)
Industry FE	Yes	Yes	Yes	Yes
N	1463	1463	1463	1463

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3 reports the results for the regression of the log of gross output on the log of each factor intensity. Each column reports the result of the regression for the specified factor intensity, where industry fixed effects are included to adjust for differences in factor intisities across industries. A value greater than zero indicates that larger firms are more likely to operate technologies with a greater factor intensity for the specified input.

Table 4

Percentage Reduction in Standard Deviations of Logged MRPs

ISIC	Capital	Skilled Labor	Unskilled Labor	Intermediate Inputs
1511	34.7	3.0	10.9	3.9
1512	40.2	20.9	9.3	-6.4
1552	37.4	16.0	34.9	-0.5
1711	41.9	15.2	35.4	8.5
2010	26.4	-0.2	18.0	2.4
2102	31.4	17.3	33.6	15.4
2221	30.7	39.9	47.3	0.6
2411	40.4	25.9	29.1	1.3
2424	28.0	17.5	7.9	3.3
2429	26.9	10.5	6.7	4.3
2520	31.0	0.8	8.8	-1.5
2695	30.7	11.9	7.4	3.0
2720	-2.4	21.5	37.0	-2.1
2811	36.6	11.2	44.4	0.8
Mean	31.0	15.1	23.6	2.4
Aggregate	15.0	6.8	12.1	1.3

Table 4 reports the percentage reduction in the standard deviation of logged MRPs of each input after accounting for the presence of multiple production technologies. Only industries with more than one production technology are included in the table. The row labeled *Mean* reports the average reduction taken across the industries listed in the table. The row labeled *Aggregate* reports the reduction when we compute the standard devition in logged MRPs across all plants in the economy while ignoring industries. Columns 2-5 show the percentage decline in the standard deviation of the logged marginal revenue product for capital, skilled labor, unskilled labor, and intermediate inputs, respectively, in each industry. The percentage decline of the standard deviation for input i is calculated as $100 * (\sigma_{i,\text{multi}}/\sigma_{i,\text{1tech}} - 1)$, where $\sigma_{i,\text{multi}}$ is the standard deviation of the logged MRP for input i allowing for multiple technologies and $\sigma_{i,\text{1tech}}$ is the same statistic for the case in which we only allow for a single production technology in each industry.

Table 5

Percentage Increase in Output from the Elimination of Misallocation

ISIC	Multiple Technologies	One Technology
1511	15.3	16.8
1512	47.3	32.5
1513	25.0	21.9
1520	17.4	16.3
1531	18.9	20.4
1541	36.7	31.3
1711	-6.9	-3.2
1810	5.5	2.0
1920	9.3	3.5
2010	2.8	6.6
2102	-8.1	-7.7
2211	-6.4	-14.3
2221	-4.6	6.5
2411	27.3	55.8
2424	21.1	10.1
2429	16.5	18.3
2520	1.5	-1.8
2695	43.3	46.6
2720	40.9	58.2
2811	1.1	2.9
2899	6.2	-1.6
3610	9.4	0.7
Aggregate	26.8	38.7

Table 5 reports the percentage increase in industry-level output from removing distortions to eliminate misallocation both within and across industries. Column *ISIC* reports the industry code for each of the industries included in the exercise. ISIC 1552 (Manufacture of wines) is excluded because of a negative implied markup when only a single production technology is used. *Multiple Technologies* lists the calculated increase in output when we allow for multiple production technologies. *One Technology* reports the same calculation when we only allow for a single production technology. The row *Aggregate* reports the overall gain in output, where overall output is calculated according to equation 30.

Table 6

TFP Gains from Reallocation
One vs. Multiple Technologies

ISIC	Multiple Technologies	One Technology	Percentage Change
1511	27.0	32.5	-17.0
1512	34.4	38.2	-9.9
1513	33.1	33.1	.
1520	21.2	21.2	.
1531	29.7	29.7	.
1541	28.0	28.0	.
1711	9.4	18.1	-48.3
1810	20.1	25.3	-20.2
1920	14.1	14.1	.
2010	16.5	24.1	-31.4
2102	7.8	13.9	-43.8
2211	14.6	15.3	-4.5
2221	6.8	36.2	-81.2
2411	10.3	17.4	-40.6
2424	10.7	13.6	-21.3
2429	22.7	30.9	-26.7
2520	15.3	16.3	-6.4
2695	30.7	37.3	-17.7
2720	42.3	51.9	-18.5
2811	19.6	26.5	-26.0
2899	27.2	27.2	.
3610	23.0	23.0	.
Aggregate	24.3	30.9	-21.3

Table 6 reports the percentage increase in industry-level TFP from removing distortions to eliminate misallocation both within and across industries. Column *ISIC* reports the industry code for each of the industries included in the exercise. ISIC 1552 (Manufacture of Wines) is excluded because of a negative implied markup when only a single production technology is used. *Multiple Technologies* lists the calculated increase in output when we allow for multiple production technologies. *One Technology* reports the same calculation when we only allow for a single production technology. *Percentage Change* reports the percentage difference between the predicted TFP gains from eliminating misallocation under multiple production technology compared to the assumption of a single production technology. The row *Aggregate* reports the overall gain in output, where overall output is calculated according to equation 30.

References

Alfaro, L., A. Charlton, and F. Kanczuk (2009): "Plant-Size Distribution and Cross-Country Income Differences," in *NBER International Seminar on Macroeconomics 2008*, ed. by J. Frankel, and C. Pissarides, pp. 243–272. Chicago: University of Chicago Press.

Alvarez, R., and R. A. López (2005): "Exporting and Performance: Evidence from Chilean Plants," *Canadian Journal of Economics/Revue canadienne d'économique*, 38(4), 1384–1400.

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. Society for Industrial and Applied Mathematics.

Asker, J., A. Collard-Wexler, and J. De Loecker (2014): "Dynamic Inputs and Resource (Mis)Allocation," *Journal of Political Economy*, 122(5), 1013–1063.

Asturias, J., M. García-Santana, and R. Ramos (2016): "Competition and the Welfare Gains from Transportation Infrastructure: Evidence from the Golden Quadrilateral of India," *CEPR Discussion Paper DP11283*.

Asturias, J., S. Hur, T. J. Kehoe, and K. J. Ruhl (2017): "Firm Entry and Exit and Aggregate Growth," *NBER Working Paper 23202*.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2013): "Cross-Country Differences in Productivity: The Role of Allocation and Selection," *American Economic Review*, 103(1), 305–334.

Bergoeing, R., and A. Repetto (2006): "Micro Efficiency and Aggregate Growth in Chile," *Cuadernos de Economia*, 43(127), 169–191.

Bils, M., P. J. Klenow, and C. Ruane (2017): "Misallocation or Mismeasurement," *Working Paper*.

Busso, M., L. Madrigal, and C. Pagés (2013): "Productivity and Resource Misallocation in Latin America," *B.E. Journal of Macroeconomics*, 13(1), 903–932.

Calinski, T., and J. Harabasz (1974): "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, 3(1), 1–27.

Crone, T. M. (2005): "An Alternative Definition of Economic Regions in the United States Based on Similarities in State Business Cycles," *Review of Economics and Statistics*, 87(4), 617–626.

Dias, D. A., C. R. Marques, and C. Richmond (2016): "Misallocation and Productivity in the Lead Up to the Eurozone Crisis," *Journal of Macroeconomics*, 49, 46–70.

García-Santana, M., E. Moral-Benito, J. Pijoan-Mas, and R. Ramos (2016): "Growing Like Spain: 1995-2007," *Working Paper*.

Gartner, W. B. (1990): "What Are We Talking about When We Talk about Entrepreneurship?," *Journal of Business Venturing*, 5(1), 15–28.

Gopinath, G., S. Kalemli-Ozcan, L. Karabarbounis, and C. Villegas-Sanchez (2015): "Capital Allocation and Productivity in South Europe," *NBER Working Paper 21453*.

Guner, N., G. Ventura, and Y. Xu (2008): "Macroeconomic Implications of Size-Dependent Policies," *Review of Economic Dynamics*, 11(4), 721–744.

Hopenhayn, H. A. (2014): "Firms, Misallocation, and Aggregate Productivity: A Review," *Annual Review of Economics*, 6(1), 735–770.

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

Humphries, J. E. (2017): "The Causes and Consequences of Self-Employment over the Life Cycle," *Working Paper*.

Jain, A. K. (2010): "Data Clustering: 50 Years Beyond K-means," *Pattern Recognition Letters*, 31(8), 651–666.

Jakobsson, M., and N. A. Rosenberg (2007): "CLUMPP: A Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure," *Bioinformatics*, 23(14), 1801–1806.

Kotey, B., and G. G. Meredith (1997): "Relationships among Owner/Manager Personal Values, Business Strategies, and Enterprise Performance," *Journal of Small Business Management; Milwaukee*, 35(2), 37–64.

Leuz, C., D. Nanda, and P. D. Wysocki (2003): "Earnings Management and Investor Protection: An International Comparison," *Journal of Financial Economics*, 69(3), 505–527.

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," *Review of Economic Studies*, 70(2), 317–341.

Levy-Yeyati, E., and F. Sturzenegger (2005): “Classifying Exchange Rate Regimes: Deeds vs. Words,” *European Economic Review*, 49(6), 1603–1635.

Li, K. (2015): “Privatization, Distortions, and Productivity,” *Working Paper*.

Oberfield, E. (2013): “Productivity and Misallocation During a Crisis: Evidence from the Chilean Crisis of 1982,” *Review of Economic Dynamics*, 16(1), 100–119.

Park, S. H., I. D. Yun, and S. U. Lee (1998): “Color Image Segmentation Based on 3-D Clustering: Morphological Approach,” *Pattern Recognition*, 31(8), 1061–1076.

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

Peters, M. (2013): “Heterogeneous Mark-Ups, Growth, and Endogenous Misallocation,” *Working Paper*.

Petrin, A., and J. Levinsohn (2012): “Measuring Aggregate Productivity Growth Using Plant-level Data,” *RAND Journal of Economics*, 43(4), 705–725.

Petrin, A., and J. Sivadasan (2013): “Estimating Lost Output from Allocative Inefficiency, with an Application to Chile and Firing Costs,” *Review of Economics and Statistics*, 95(1), 286–301.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000): “Inference of Population Structure Using Multilocus Genotype Data,” *Genetics*, 155(2), 945–959.

Ray, S., and R. H. Turi (1999): “Determination of Number of Clusters in K-Means Clustering and Application in Colour Segmentation,” in *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pp. 137–143. Calcutta, India.

Restuccia, D., and R. Rogerson (2008): “Policy Distorsions and Aggregate Productivity with Heterogeneous Establishments,” *Review of Economic Dynamics*, 11(4), 707–720.

Sandleris, G., and M. L. J. Wright (2014): “The Costs of Financial Crises: Resource Misallocation, Productivity, and Welfare in the 2001 Argentine Crisis,” *Scandinavian Journal of Economics*, 116(1), 87–127.

Schelkle, T. (2016): “General Methods for Measuring Factor Misallocation,” *Working Paper*.

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

Ziebarth, N. L. (2013): “Are China and India Backward? Evidence from the 19th Century U.S. Census of Manufactures,” *Review of Economic Dynamics*, 16(1), 86–99.

A Appendix

A.1 Monte Carlo Exercise

This section evaluates the performance of our clustering methodology using simulations. We focus on our ability to identify when multiple production technologies are used within an industry and to correctly identify which firms use each production technology. We find our methodology performs well in recovering unobserved production technologies when data is generated according to the economic model in Section 2 and the assumption in equation 14 holds. In order to explore the robustness of our methodology we also evaluate our methodology under a variety of sample sizes and methodological choices including those in which the assumption in equation 14 is violated.

A.1.1 Simulation Strategy

This section reports the process we use to generate the simulated data that we use to evaluate our methodology. For each simulation we generate a single industry which we assign to have either one, two, or three production technologies with equal probability. Each production technology is defined as a Cobb-Douglas production function with four inputs and varying factor intensities.

In the data, there are frequently one or two factors intensities much larger than the others, therefore, we choose the factor intensities in a two step process. For each production technology we randomly choose the number of inputs that the technology will be more-intensive in, where the technology can be more-intensive in either one, two or three inputs, chosen with equal probability. Conditional on a given production technology being more-intensive in a specific number of inputs, we randomly assign which factor inputs are the more-intensive inputs with equal probability and we assign the remaining factor inputs as less-intensive. In order to ensure the production technologies are reasonably distinct and thereby satisfy the assumption in equation 14 we require that no two production technologies in the industry are identical in terms of what their more-intensive and less-intensive factor inputs are. For factors classified as more-intensive we assign them a preliminary value drawn from a uniform distribution with support of [0.20, 1.00], whereas factors classified as less-intensive have a preliminary value drawn from a uniform distribution with support of [0.01,0.10]. This leads to a median ratio between the largest and smallest factor intensities of 22, compared to 23 in the data. We then calculate the factor intensities, θ_h^i, for each production technology, h, by normalizing the preliminary values to sum to 1.

For each production technology, we assign a random number of firms to use the technology, where we use a uniform distribution to generate between 30 and 70 firms for each technology. This ensures we have enough firms to apply our methodology for each simulation. For each firm, m, we generate the distortions, τ_m^i, for each input, i, drawn from a multivariate lognormal distribution. We set the parameters so the mean of the logged distribution is zero for all inputs, and the the standard deviations are chosen from a uniform distribution with support [0.1, 0.3]. Using

Table A.1

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: [30,70]
Standard deviation of lognormal distribution of input i distortion	Uniform: [0.1,0.3]
Technology specific distortion (τ^h)	Uniform [0.5,1.5] (re-scaled to mean of 1)
Parameter that governs elasticity of substitution (ρ)	Markup ($1/\rho$) is uniform: [1.01,1.50]

Table A.1 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.

a multivariate distribution allows for the possibility of correlation between the distortions. In the base scenario we assume the distortions are independently distributed, however, we evaluate correlated distortions in Table A.5. The average correlation in the data is close to zero, however, the standard deviations tend to be larger than the standard deviations we use in our simulations. We chose the standard deviations to be smaller to ensure that the assumption in equation 14 holds, however, we later explore the robustness of our results to generating distortions from distributions with larger standard deviations.

In order to apply our clustering methodology to our simulated data, we need to calculate the cost shares for each firm using equation 8. When calculating these shares, we require an elasticity of substitution and the technology-specific distortions. The median markup in the data is 1.25, with a range between 1.0 and 1.5. Therefore, we select the elasticity of substitutions so that markups are uniformly distributed across simulations, ranging between a 1 percent and 50 percent markup. The technology specific distortion is chosen by drawing a preliminary value from a standard uniform distribution for each production technology, and then normalizing the preliminary values so that the mean value is 1. Because our methodology only requires cost shares, and not revenues and expenditures separately, other parameters such as the the firm-level productivity, z_m, of each firm are irrelevant for our evaluation.

Table A.1 summarizes the parameters used for the Monte Carlo exercise.

Table A.2

True vs. Recovered Number of Production Technologies

	Number of Technologies Recovered			Percent
True Number of Technologies	1	2	3	Correct
1	100.0	0.0	0.0	100.0
2	0.0	100.0	0.0	100.0
3	0.0	0.0	100.0	100.0

Table A.2 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 A.3

Percentage of Firms Correctly Classified by Technology

	True Number of Technologies			
	1	2	3	Overall
Average Percentage	100.0	100.0	100.0	100.0

Table A.3 shows the percent of firms that were correctly classified by their technology use in the Monte Carlo exercise. The columns labeled *1, 2,* and *3*, indicate the number of production technologies in the simulated data, while the column *Overall* indicates the average across all simulations.

A.1.2 Simulation Results

After generating the synthetic data, we run all of the steps of the clustering methodology described in Section 3. Table A.2 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 100 percent of the time regardless of whether there are 1, 2, or 3 technologies. Table A.3 shows the percentage of firms that we correctly classify by technology, which is also 100 percent.

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:

$$\text{Absolute Percentage Error} \equiv \left| \frac{\Omega^{\text{true}} - \Omega^{\text{recovered}}}{\Omega^{\text{true}}} \right|. \tag{43}$$

The results for the factor intensities, distortions, and markups are reported below in Table A.4. For all variables, the average absolute percentage error is less than 10 percent, indicating a high degree of accuracy in parameter recovery. For example, the value for factor intensities is around 3 percent, indicating that if the true factor intensity is 0.35, our results suggest that the recovered

Table A.4

Average Absolute Percent Error between True and Recovered Parameters

	True Number of Technologies			
	1	2	3	Overall
Factor Intensities (θ_h^i)	2.7	2.8	2.7	2.7
Idiosyncratic Distortions (τ_m^i)	4.6	4.6	4.4	4.5
Tech-Specific Distortions (τ_h)	0.0	4.9	6.5	3.8
markup	4.2	8.0	9.5	7.2

Table A.4 shows the Absolute Percentage Error (as described in equation 43) 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.

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.

In order to explore the robustness of our methodology, we report the success rate of our methodology under two additional scenarios. Table A.5 reports the percentage of simulations in which we correctly recover the true number of production technologies for each scenarios. Base refers to our base case scenario that we describe in Section A.1.1. The other columns of the table report deviations from the base scenario as follows. Correlated Taus refers to the case where we allow the distortions to be correlated across factor inputs, where the correlations are drawn randomly from a uniform distribution with support $[-0.2, 0.2]$. Overlapping Clusters refers to the case where we increase the standard deviation of the distortions — we generate them from a uniform distribution with support $[0.3, 0.7]$ compared to the base scenario support of $[0.1, 0.3]$ — which leads to the assumption in equation 14 being frequently violated. Under both scenarios we find that our methodology performs well overall, finding the true number of technologies in 99.3 percent of simulations with correlated distortions and 93.7 percent of simulations with overlapping clusters. When the clusters overlap, however, we find we are more likely to understate the true number of clusters compared to the base scenario. This result indicates the conservative nature of our methodology, and the inherant difficulty of disentangling closely related technologies.

Table A.5

Robustness of Methodology in Alternative Scenarios

	True Number of Technologies			
Number of Technologies	1	2	3	Overall
Base Scenario	100.0	100.0	100.0	100.0
Correlated Distortions	99.0	100.0	100.0	99.7
Overlapping Clusters	100.0	100.0	81.0	93.7

Table A.5 reports the success rate of our methodology — the percentage of simulations in which we successfully uncover the true number of production technologies — under three scenarios. The first column reports the scenario we are using to generate our simulations. Columns 2-4 report the success rates for simulations for each number of production technologies. The column *Overall* indicates the success rate across all simulations.