

Understanding the Foundations of Product Scope

Matthew Flagge

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy  
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016



## ABSTRACT

### Understanding the Foundations of Product Scope

Matthew Flagge

The following essays examine the nature of product co-production patterns in India—what factors cause these patterns to emerge, and why they are valuable to study. The first chapter establishes a motivation. It takes a measure of product co-production established in the literature—the “proximity” matrix of Hidalgo et al. (2007)—and shows that this measure is an excellent predictor of new product additions by firms and states, even controlling for other potentially relevant explanatory variables. The following chapter employs a reduced-form approach with regression analysis to uncover the factors that could be giving rise to these patterns of co-production. Using this approach, demand complementarities and patterns of input similarity seem to have the most explanatory power for the observed patterns.

The final chapter improves the estimation by incorporating product paths and firm profitability into a structural model. We adapt the gravity model of Morales et al. (2015) to our setting to identify costs associated with adding new products based on characteristics of the relationship between the firm and its potential products. In the model, firms seek to expand their product scope into the most profitable products, where this profit is diminished by “distance” they would have to traverse through a characteristic space. Using the moment inequalities method of Pakes et al. (forthcoming), we are able to estimate which dimensions in that space have the greatest effect on firm profits. We find the physical distance to the nearest location of production had the greatest impact, followed by input similarity between their products and potential products.

# Table of Contents

List of Charts, Graphs, and Illustrations.....	iv
Chapter 1: Predictive Power of Proximity on Firm- and State-level production in India .....	1
1    Introduction .....	2
2    Data .....	4
3    Computing Proximity .....	6
3.1    Revealed Comparative Advantage (RCA).....	7
3.2    Proximity .....	7
4    Predictive Power of Proximity .....	9
4.1    Conditional Probability of Adding and Proximity of Nearest Product .....	10
4.2    Proximity Density (Average Proximity) .....	12
4.3    Logit Regressions.....	15
5    Conclusion.....	23
6    References .....	25
7    Appendix .....	27
Chapter 2: Foundations of Product Proximity .....	36
1    Introduction .....	37
2    Data.....	39
3    Product-to-Product Relationships.....	40
3.1    Output Proximity.....	41
3.2    Input Proximity (Demand Complementarities).....	43
3.3    Input Similarity (Cost Complementarities).....	44
3.4    Labor Intensity .....	45
3.5    Urban Share .....	47
3.6    Firm Size .....	47
4    Results.....	48
4.1    Variance Partitioning.....	58
5    Conclusion.....	60
6    References .....	62
Chapter 3: Firm-Product Linkages and the Evolution of Product Scope.....	64

1	Introduction .....	65
2	Data .....	69
3	Preliminary Evidence.....	72
4	Theoretical Framework.....	74
4.1	Demand.....	74
4.2	Supply.....	75
4.2.1	Marginal Costs.....	75
4.2.2	Fixed Costs .....	76
4.2.3	Product Startup Costs .....	77
4.3	Firms' Optimal Behavior .....	80
5	Estimation .....	85
5.1	First Stage.....	85
5.2	Second Stage .....	87
5.2.1	Selecting Observations for Moments .....	88
5.2.2	Aggregating Observations into Moments.....	94
5.2.3	Estimating the Bounds .....	95
5.2.4	Properties of the Error Terms .....	96
5.2.5	Confidence Intervals .....	97
6	Results.....	97
7	Supplementary Analyses.....	103
7.1	Number of Profitable Products .....	105
7.2	Firms' Product Choices.....	106
7.3	Network Density Regression .....	109
8	Conclusion.....	111
9	References .....	113
10	Appendix .....	116
10.1	Marginal Cost Regression.....	116
10.1.1	<i>PLj</i> .....	116
10.1.2	<i>PICfjt</i> (Intermediate Input Costs) .....	118
10.1.3	Input-Output Table .....	118

10.1.4	Assigning Inputs to Outputs at the firm level .....	120
10.1.5	Regression Results .....	123
10.2	Physical Distance Calculation .....	123
10.2.1	Mapping firms to districts .....	123
10.2.2	Calculating the Distance .....	124
10.3	Alternative Specifications .....	126
10.3.1	Large Firms Only .....	126
10.4	Moments for Baseline specification .....	128

# List of Charts, Graphs, and Illustrations

---

Figure 1 – Conditional Probability of Adding Products (State) .....	11
Figure 2 – Conditional Probability of Adding Product (Firm).....	12
Figure 3 – Density of Productive Structure around Transition vs. Undeveloped Products (State) .....	13
Figure 4 – Density of Productive Structure around Transition vs. Undeveloped Product (Firm) .....	14
Figure 5 - Conditional Probability of Adding Products (State).....	28
Figure 6 - Conditional Probability of Adding Products (Firm) .....	29
Figure 7 - Density of Productive Structure around Transition vs. Undeveloped Product (State).....	30
Figure 8 - Density of Productive Structure around Transition vs. Undeveloped Product (Firm).....	31
Figure 9 – Cross-Section of the Estimated Set for Animals, Vegetables, and Forestry.....	100
Figure 10 – Cross-Section of the Estimated Set for Rubber, Plastic, and Leather .....	100
Figure 11 – Cross-Section of the Estimated Set for Textiles .....	101

## Acknowledgements

I would like to thank Eric Verhoogen, Amit Khandelwal, Eduardo Morales, David Weinstein, Rich Clarida, Kate Ho, Jon Vogel, Pete Schott, Don Davis, Peter Neary, Jonathan Dingel and Chris Conlon for helpful comments, conversations, and feedback during the process of writing this dissertation. I also express my gratitude to Eric and Amit for facilitating access to the ASI data. The author accepts responsibility for any errors.



# Dedication

Dedicated to my Father and Mother.

# Chapter 1: Predictive Power of Proximity on Firm- and State-level production in India\*

Matthew Flagge<sup>1</sup>

<sup>1</sup>Columbia University

June 2015

I compute proximities between products in India based on co-production correlations following the formula devised by Hidalgo et al. (2007). I perform several tests to determine if our Indian proximity matrix is informative for predicting the development of new products at the firm and state level within India. I look at the proximity of the closest product in a producer's basket to potential products, as well as the density of a producer's basket around potential products. All tests confirm the importance of these proximities for predicting the development of new products within states and firms.

---

\* I would like to thank Eric Verhoogen and Amit Khandelwal for helpful comments and for facilitating access to the Annual Survey of Industries data. The author accepts responsibility for any errors.

# 1 Introduction

Can the presence of related industries predict which new products a firm starts producing or a country starts exporting? Much of the study of international trade has focused on why countries develop comparative advantage in some products as opposed to others. Classical literature on the subject (Heckscher and Ohlin 1991) has focused on differences in relative factor intensities across countries, but this theory has been shown to be unsatisfactory for explaining the tendency of certain clusters of products to be exported simultaneously (Leamer 1984).

Hidalgo et al. (2007) have shown that countries have a tendency to move towards products in their current clusters. That is, if two products are typically exported together, a country that has comparative advantage in only one of the two products is likely to develop comparative advantage in the other. Hidalgo and co-authors are agnostic about the causes of this phenomenon, but numerous theories in both classic and more recent literature put forth reasons why this might be true.

Marshall (1920) argued that certain industries could create positive externalities for other related industries in their area in the form of knowledge spillovers, shared labor pools, or reduced transportation costs for inputs or outputs. Frenken and Boschma (2007) and Winter (2006) argue that technological progress often takes the form of incremental improvements over existing technology, leading firms to expand into products with production technologies similar to those they already have. Maskell and Malmberg (2007) argue that micro-level myopic search as well as macro-level institutions constrain the set of innovations firms and entrepreneurs undertake, leading to clusters. Our third chapter in this dissertation (Flagge and Chaurey, 2015) also puts forth a model wherein firms must pay a higher cost to initiate production in products more dissimilar to their existing basket.

These theories suggest that expansion into related products should be observed in firms' choices of products to produce, as well as countries' export baskets. Hidalgo et al.'s measure of proximity between products has been shown to predict comparative advantage in exports at the level of

countries, regions, and firms (Poncet and de Waldemar,[2015]; Boschma et al. [2013]). Neffke and Henning (2008) develop a measure similar to Hidalgo et al.'s to capture co-occurrence of different industries within the same firm using production data, and they find their measure of industry relatedness is useful for explaining how regions diversify into new industries. In a related study, Colombelli et al. (2014) employ Hidalgo et al.'s (2007) methodology to build a matrix of proximities between technologies using patent data and find it to be informative for explaining regions' expansions within the nanotechnology sector.

However, we are unaware of any study testing the relevance of Hidalgo et al.'s proximity measure for product choice at the firm level. This paper uses their formulae to compute a proximity matrix measuring co-production correlations between products in India and tests its power for predicting new products added by firms. Consistent with the established theory on firm expansion, the results find that existing co-production correlations are informative for predicting future firm product choice. However, with a reduced form model, it is impossible to know which firms will add products, since our current setting gives us no insight into which firms are sufficiently profitable to allow the addition of products, how much it costs to add new products (and whether those costs would be justified by the expected profitability of the products), or what underlying features of product proximity are most relevant to reducing those costs. For this, we will need a structural model incorporating profits and costs, which will be presented in our third chapter (Flagge and Chaurey, 2015).

The data we use come from India's Annual Survey of Industries (ASI). This is a factory-level dataset with yearly observations that is a representative sample of manufacturers in India. It includes disaggregated (5-digit) data on products produced in each factory during a time of rapid expansion of firms' product scope.

We find that even controlling for a number of firm characteristics, Hidalgo et al.'s proximity measure seems to be an excellent predictor of which products will be added by firms. In logit regressions of the addition of new products on a vector of controls, both proximity to the nearest developed product in a producer's basket, as well as average proximity of potential products to producers' baskets, turn out to be better predictors of new product additions than any of our available firm or state characteristics. The result holds at both the state and the firm level, and is robust to whether or not we considered producers simply adding products, or if they are actually required to develop a comparative advantage in those products (relative to other producers in India).

The paper is organized as follows. Section 2 describes the data. Section 3 discusses how Hidalgo et al. (2007) computed their proximity matrix, and how we do it in this paper. Section 4 goes through several tests of the significance of the proximity matrix in India, and compares them to the results received by Hidalgo et al. (2007). Section 5 concludes.

## **2 Data**

As mentioned in the introduction, our primary dataset is the Annual Survey of Industries (ASI) from India. It is an unbalanced factory-level panel with yearly observations spanning 1999 – 2008, and presents a representative sample of all factories that have 10 or more employees with power, or 20 or more employees without power.

For this study, we make use of the section of the survey that reports factory outputs. We observe both the sale value of all outputs, as well as the 5-digit product category of those outputs. The data are fairly disaggregated, with 5204 products available at the 5-digit level and 262 products at the 3-digit level.

In this paper we will be examining the predictive power of product proximity at both the state and the factory level. The ASI panel provides the state in which each of the factories is located. After dropping all factories that report only the sum of their outputs (instead of specifying individual product categories) as well as those which did not fill in the needed survey blocks for the baseline year of our specification, our sample has 31 states<sup>1</sup> and 33,871 factories represented.

It is also worth noting that 95.5% of the firm-year observations<sup>2</sup> in our data correspond to single-factory firms. This is important because one might imagine that expansion strategies and co-production correlations within factories might be different for a single-factory firm, as opposed to a single factory within a multi-factory firm. The prevalence of single-factory firms in our data indicates that this should not be of much concern for our analysis, and we will use factory and firm interchangeably for the remainder of the paper. Other summary statistics for the ASI data are found in Table 1.

The other data we used were the revealed comparative advantage and proximity calculations<sup>3</sup> produced by Hidalgo et al. (2007), and available in their online appendix.<sup>4</sup> These were, in turn, computed from Feenstra et al.'s (2005) World Trade Flows dataset, available from the NBER.

---

<sup>1</sup> Readers familiar with the geography of India might observe that it only has 29 states. The reason for this discrepancy is that our sample excludes the small states of Arunachal Pradesh, Mizoram, and Sikkim. The modern state of Telangana was not formed until 2014, which was after our sample period, but was included in our data under the state of Andhra Pradesh. Our data includes the non-state union territories of the Andaman and Nicobar Islands, Chandigarh, Dadra and Nagar Haveli, Daman and Diu, Delhi, and Pondicherry, all of which we treat equivalently to states for the purposes of this analysis, and we will informally refer to them as states throughout the paper.

<sup>2</sup> There are 347,857 firm-year observations in the data.

<sup>3</sup> We will explain shortly what both revealed comparative advantage and proximity are.

<sup>4</sup> <http://www.chidalgo.com/productspace/data.htm>

**Table 1 – Mean values of Producer Characteristics**

	<b>Firms</b>	<b>States</b>
Urban	0.597	0.532
Fixed Assets	112.44 (1654)	79.48 (66.36)
Working Capital	20.35 (364.64)	18.14 (12.94)
Average number of workers	160.32 (702.45)	117.53 (60.77)
Total wages	12.42 (90.33)	9.18 (7.31)
Expenses	17.38 (114.43)	12.57 (9.61)
Observations	33871	31

*Notes: Standard errors in parentheses. Fixed assets, working capital, total wages, and total expenses are given in millions of rupees. Statistics are for the year 2003. Fixed assets include plant & machinery, land, buildings, equipment, etc. Working capital includes raw materials, cash-on-hand, finished goods, etc.. Expenses includes repair and maintenance, operating expenses, rent, etc. State statistics are based on the average across all the firms within that state.*

### **3 Computing Proximity**

The measure of product “closeness,” or proximity used by Hidalgo et al. (2007) is based on revealed comparative advantage (RCA) in exports. Intuitively, they define “proximity” between two products as the conditional probability of having RCA in one product while simultaneously having RCA in the other product.

### 3.1 Revealed Comparative Advantage (RCA)

The notion of revealed comparative advantage was developed by Balassa (1965). It is based on the idea that since relative costs and other factors that influence trade are often difficult to observe, the pattern of trade itself might reveal the comparative advantage of a particular exporter.

Balassa's method of doing this involved comparing the relative shares of countries' exports of particular products. He normalized these numbers by dividing by that country's share in the total exports of all commodities from all countries under consideration. Thus, the RCA of an exporter  $e$  for a particular product  $p$  is given by:

$$RCA(e, p) = \frac{\frac{x_{ep}}{\sum_p x_{ep}}}{\frac{\sum_e x_{ep}}{\sum_{ep} x_{ep}}} \quad (1)$$

Where  $x_{ep}$  is the total exports of product  $p$  by exporter  $e$ . The result is an index that tells whether a particular producer exports more ( $RCA > 1$ ) or less ( $RCA < 1$ ) of a product than average. Put a different way, if a particular exporter  $e$  has an RCA of 1.1 for product  $p$ , that means its share of the exports of  $p$  is 10% higher than its share of the total exports in the sample, it we would say that  $e$  has a revealed comparative advantage in  $p$ .

In this paper, because we are using factory-level production data, rather than export data, we calculate the RCA measure based on outputs rather than exports. The same formula applies, except now the denominator in the equation is based on the total production of India, rather than total exports for the world.

### 3.2 Proximity

The proximity measure we use in this paper follows the definition used by Hidalgo et al. (2007). It is based on the idea that a country that has the necessary resources and capabilities to produce one



product will also be able to leverage those inputs to the production of other similar products. Rather than try to specify what those inputs might be, they rely on an output-based measure, and assume that if two products are frequently produced together, they must therefore require similar inputs.<sup>5</sup>

To measure the proximity between two products, they calculate the probability of having  $RCA > 1$  in one product, conditional on having  $RCA > 1$  in the other product. Then, because they want their proximity measure to be symmetric (just as a distance measure would be symmetric), they define the proximity as the minimum of these two conditional probabilities. Thus, the proximity  $\phi^1$  between products  $p$  and  $q$  would be:<sup>6</sup>

$$\phi^1(p, q) = \min\{P(RCA(p) > 1 | RCA(q) > 1), P(RCA(q) > 1 | RCA(p) > 1)\} \quad (2)$$

This proximity measure can be calculated for every pair of products. The full matrix of proximities for every product pair has come to be known as the Product Space.

Our use of production data inside the boundaries of a country allows us to see which products are being produced together much more clearly than Hidalgo et al. were able to see with the World Trade Flows data. Because we observe firms rather than countries, the smaller scale of the producers in our data also makes it less likely that any individual agent produces *every* product. We therefore consider two possible proximity measures in the present study. The first is the measure of Hidalgo et al., given in equation (2). The second is a measure based on whether two products are produced together at all (as opposed to requiring the producers to exhibit revealed comparative advantage over

---

<sup>5</sup> Here, we should understand inputs to mean not only capital, labor, or even intermediate inputs, but rather the whole array of knowledge, institutions, trading connections, local resources, and other factors necessary to produce a product.

<sup>6</sup> Note that we are using the superscript on the  $\phi$  to indicate the cutoff value for the RCA.

the rest of India). In practice, this simply means having  $RCA > 0$ , so our alternate proximity measure,  $\phi^0$ , can be computed as follows:

$$\phi^0(p, q) = \min\{P(RCA(p) > 0 | RCA(q) > 0), P(RCA(q) > 0 | RCA(p) > 0)\} \quad (3)$$

## 4 Predictive Power of Proximity

One of the reasons the Product Space gained attention is that it proved useful for predicting which products a country would develop a revealed comparative advantage in. Hidalgo et al. calculated the matrix of product proximities using data from the years 1998 – 2000. They then defined the set of transition products as those products for which a country had  $RCA < 0.5$  in 1990, and  $RCA > 1$  in 1995. These are the products for which a country developed a revealed comparative advantage over the observed period.

They then measured how “close” these transition products were to the countries’ previous export basket, as measured by the proximities between the products. The results, shown in their paper, speak strongly in favor of the predictive power of their proximity measure.

We perform similar exercises here, testing the predictive power of our Indian proximity matrices for the expansion of product baskets by firms and states within India. The fact that our dataset spans different years than that used by Hidalgo et al. means that we cannot use the same time window as they have in their study.<sup>7</sup> However, to stay as close as possible to their methodology, we too examine transition products over a five-year span: 2004 – 2008. We will call this span the “transition period,”

---

<sup>7</sup> Our data covers 1999 – 2008, while the World Trade Flows data used by Hidalgo et al. (2007) spans 1962 – 2000.

and the basket of products a producer had in the year *prior* to the transition period (here, 2003), we will term the “pre-transition basket.”<sup>8</sup>

As explained in the previous section, we will consider two different proximity matrices for examining product additions over the transition period. The first is based on the RCA-based proximity  $\phi^1$ , and the second is based on the production-based proximity  $\phi^0$ .<sup>9</sup> Like Hidalgo et al. (2007), we compute our final matrices by averaging over three of the year-specific proximity matrices. Hidalgo et al. used 1998 – 2000, and we will use 2001 – 2003. This allows us to generate a proximity matrix that is computed entirely with data prior to the transition period, eliminating the potential for simultaneity or reverse causality.

#### **4.1 Conditional Probability of Adding and Proximity of Nearest Product**

We will look at several possible relationships between a producer and the set of potential products. One such measure of the relevance of proximity is by computing the probability that a producer will develop revealed comparative advantage in a product, conditional on the closest product in its pre-transition basket being at proximity  $\phi$ .

Hidalgo et al. (2007) calculated these probabilities and found they were monotonically increasing in the proximity of a producer’s nearest product. That is to say, the closer one of a producer’s products is to a given product, the higher the probability that producer will develop  $RCA > 1$  in that product in the near future.

---

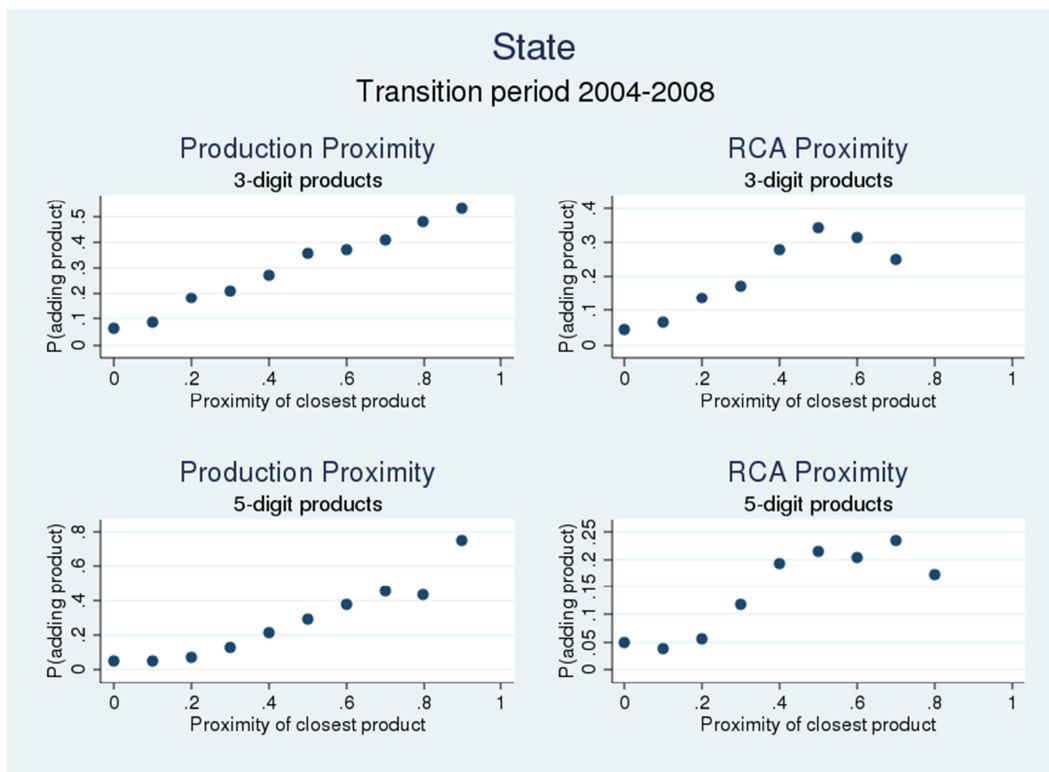
<sup>8</sup> As a robustness check, we will also include the results for an extended transition period covering the seven-year span 2002 – 2008. We do not present this as our main result due to concerns about the potential endogeneity associated with using a proximity matrix based on the years 2001 – 2003, but we will include the results in the appendix.

<sup>9</sup> See equations (2) and (3), respectively.

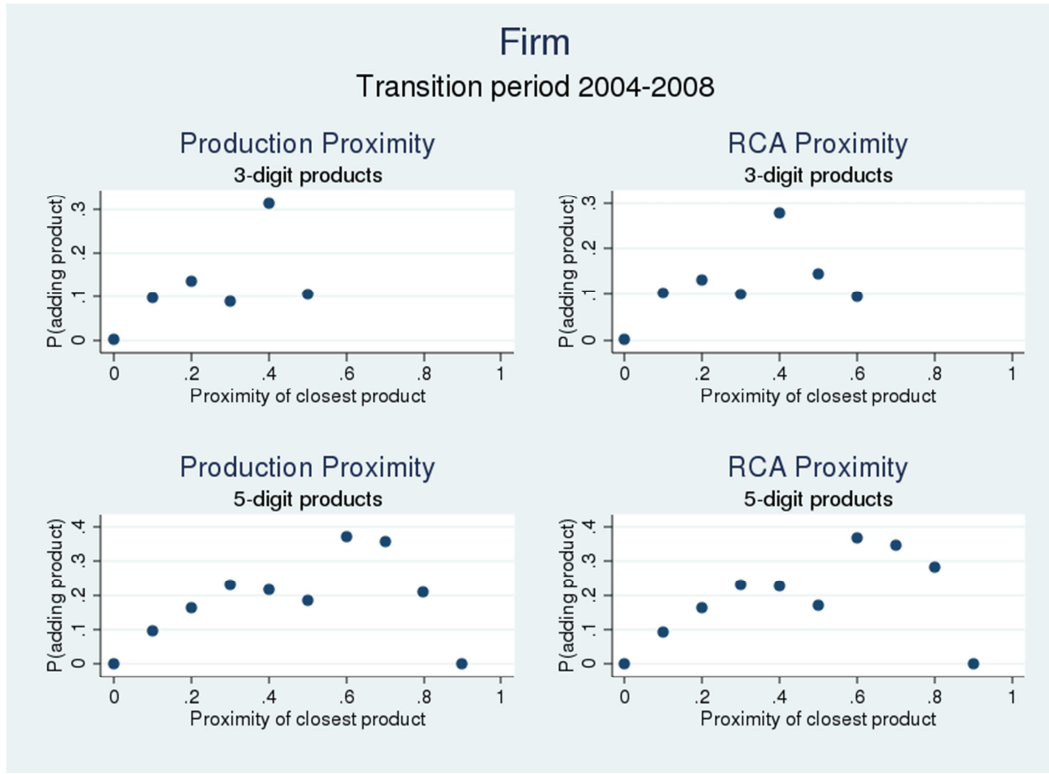
We perform the same exercise here. The results are shown in Figure 1 and Figure 2. We display the results for states and firms, for three- and five-digit product codes.

We observe that the results are not as strong as those reported by Hidalgo et al. (2007). In very few of the cases is the probability of adding a product monotonically increasing in its proximity to the state or firm's nearest product. However, this certainly is not a condemnation of their results. It appears from looking at the figures that even if the probability of adding a product is not strictly increasing in proximity, it does at least show a strong positive correlation with that measure. In the state graphs for the production-based proximity,  $\phi^0$ , the probabilities do appear to be monotonically increasing for both 3- and 5-digit products.

**Figure 1 – Conditional Probability of Adding Products (State)**



**Figure 2 – Conditional Probability of Adding Product (Firm)**



## 4.2 Proximity Density (Average Proximity)

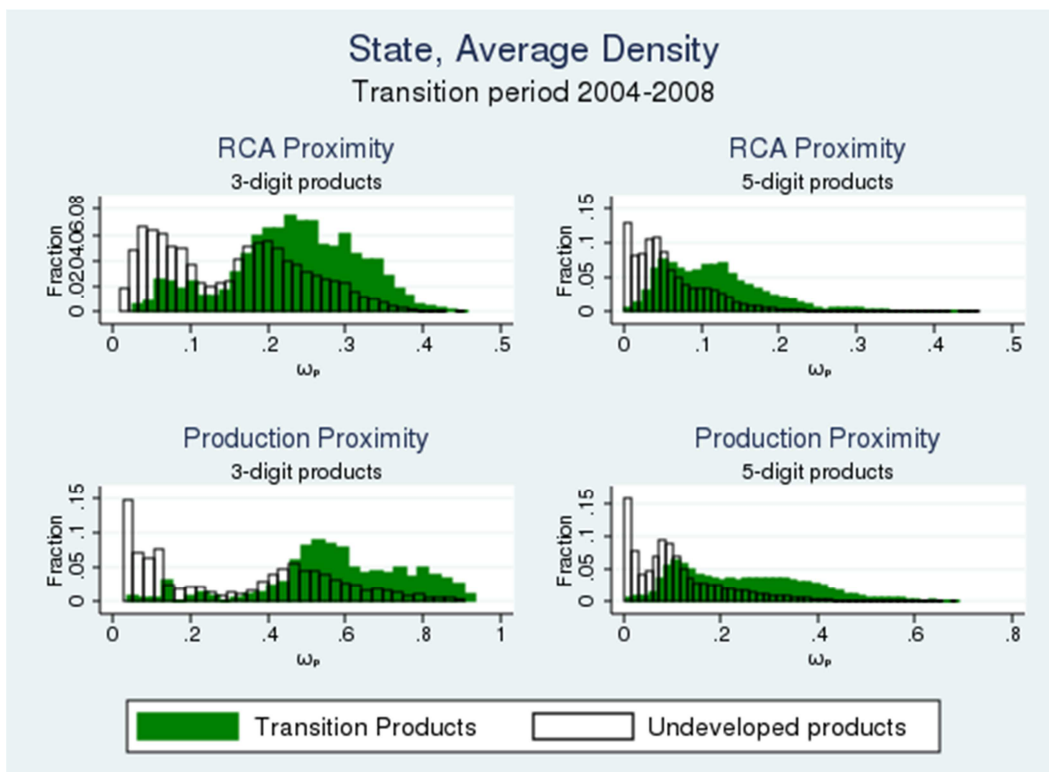
Another potential measure of the importance of proximity is not the proximity of a producer's closest product to a potential new product, but rather the average proximity of the new product to the producer's entire product basket. Hidalgo et al. (2007) termed this quantity "density," and it is a reflection of how densely the space around a particular product is populated by products already in the producer's productive structure. According to the formula used by Hidalgo et al. (2007), the density of producer  $e$ 's developed products around product  $p$  is given by:

$$\omega^x(p, q) = \frac{\sum_q r_q^e \phi^x(p, q)}{\sum_q \phi^x(p, q)} \quad (4)$$

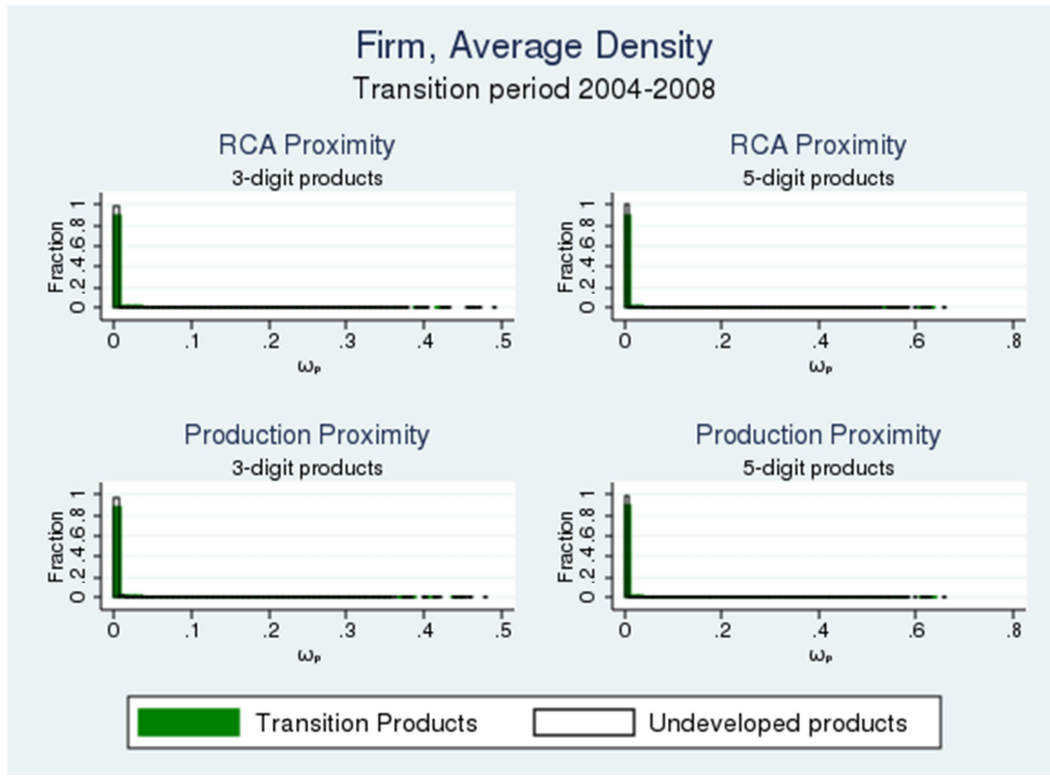
Where  $r_q^e = 1$  if producer  $e$  has  $RCA > 1$  in product  $q$  (or  $RCA > 0$  if we are considering production instead of revealed comparative advantage).

What Hidalgo et al. (2007) found is that the density of a country's productive structure around its transition products was higher on average than around the products that remained undeveloped. They used an analysis of variance (ANOVA) test to compare these distributions, and found them to be different with  $p < 10^{-30}$ . We do the same here, and the density distributions for India at the state and firm level are presented in Figure 3 and Figure 4, respectively.

**Figure 3 – Density of Productive Structure around Transition vs. Undeveloped Products (State)**



**Figure 4 – Density of Productive Structure around Transition vs. Undeveloped Product (Firm)**



It seems visually clear that the distributions for transition products vs. undeveloped products are different at the state level. This is not as clear from looking at the firm-level graphs. The firm distributions are significantly less clear. As one might expect, due to the fact the typical firm produces much fewer products than a whole state, the typical density of a firm around any product is going to be close to zero. The reason for this becomes clear upon examining equation (4)—since  $r_q^e = 0$  for almost every product for a particular firm,  $\omega(e, p) \approx 0$ .

However, following Hidalgo et al. (2007), we can confirm the difference by running ANOVA tests. The tests confirm that the distributions for the transition products and the undeveloped products are indeed different, with  $p < 10^{-170}$  for the state distributions at the 3-digit level, and  $p$  less than the

machine precision (effectively,  $p = 0$ ) for the 5-digit state distributions, and for both the 3- and 5-digit firm distributions. This confirms there are statistically significant differences between the distributions.

Given the large number of observations associated with the firm-level tests (over 425 million for the 5-digit product codes), we are hesitant to draw conclusions about the economic significance of our results, especially since a visual inspection of the distributions seems inconclusive. We acknowledge that statistical significance is almost guaranteed with so much data. However, the ANOVA tests are certainly suggestive, and the state distributions lend themselves to a clear conclusion in favor of the predictive power of the proximity matrix.

### 4.3 Logit Regressions

The evidence from the previous sections was suggestive, though admittedly not as rigorous as one might like. We included them in order to establish a baseline comparison to the results in Hidalgo et al. (2007), which included analyses of the types described in the previous sections.

One way to test the efficacy of the proximity matrix in predicting new product additions is to include the density or proximity to the closest product in a regression framework that also includes other potential explanatory variables we might expect to be correlated with the likelihood to add products. Since the dependent variable in this setting is binary (either the producer added a product or it did not), we will use a logit model, which is the standard model used in the literature for problems such as ours.

Thus, for the regressions in this section, we are estimating:

$$\Pr[RCA(e,p) > x] = \frac{\exp[Z(e,p)'\beta]}{1 + \exp[Z(e,p)'\beta]} \quad (5)$$



Where  $Z(e, p)$  represents a vector of controls which includes the variables listed in Table 3, as well as either the proximity to the closest product in producer  $e$ 's basket to product  $p$ , or the average proximity of  $e$ 's products to  $p$ . The regression will be performed only over those products for which the given producer (state or firm) did *not* have  $RCA > x$  prior to the transition period.

We should note that this is not an exercise that was performed by Hidalgo et al. (2007). Thus, to have a baseline to which we can compare, we will first show the results of such a logistic regression using their data. In this case, we are estimating the logit of a particular country developing  $RCA > 1$  for a given product  $p$ , conditional on the proximity of the closest product to  $p$  for which the country already has  $RCA > 1$ . The results are shown in Table 2.

**Table 2 – Logit New Revealed Comparative Advantage**

Regressor	Logit new RCA > 1
$\rho^1$	0.919*** (.067)
constant	-3.284*** (.023)
N	122229
Pseudo R <sup>2</sup>	0.004

*Notes: Independent variable is the proximity of the closest developed (RCA > 1) product in a country's basket to the product associated with a given observation. Observations are country-products for which a country has RCA < 1 prior to 1990, but developed RCA > 1 by 1995. Proximity matrix used is the same calculated by Hidalgo et al. (2007) by averaging proximities from the years 1998 – 2000. \*\*\* = significance at the 1% level.*

Regrettably, we do not have such rich data at the country level as we have for states and firms in India, so Table 2 displays only a univariate regression. We also have some hesitation about assigning too much weight to their results due to the possibility of reverse causality emerging from using the proximity matrix from 1998-2000 to predict countries' development of revealed comparative advantage

over the period 1990-1995. Nevertheless, we include that regression here for the sake of comparison and remind the reader that neither of these critiques will be an issue for the logistic regressions we perform on India.

In order to ensure that the proximity matrix is not merely correlated with some other aspect of a producer's ability to add products, we include a number of other potentially relevant factors based on what we have in our data. These are explained in Table 3. For each of the variables provided, the values included in the regression are the numbers for the producer in the base year (i.e. the year prior to the transition period).

**Table 3 – Explanatory Variables Included in Regressions**

Explanatory Variable	Notes	Rationale for Inclusion
Urban	1 = urban, 0 = rural	Urban producers might receive greater spillovers
Fixed Assets	Includes plant & machinery, land, buildings, equipment, etc.	Greater capital stock might enable firms to add products more easily
Working Capital	Includes raw materials, cash-on-hand, finished goods, etc..	More money or materials available to invest might facilitate product addition
Average number of workers	Total employees, including supervisors	Larger firms might be better able to add products
Total wages	Total wage bill paid to all workers	Firms with greater expenses might be more constrained in adding products
Other expenses	Includes repair and maintenance, operating expenses, rent, etc.	(see above)

*Notes: Variables for the firms are as provided in the ASI data. The equivalent variables for states are the averages across all the firms in that state for the given year.*

The results from our state-level logistic regressions are found in Table 4 and Table 5. The firm-level regressions are in Table 6 and Table 7. In the tables,  $\omega$  represents average density, as explained in section 4.2, and  $\rho$  represents the proximity to the nearest product in the producer's basket.

Of note in the tables is that the proximity measure is positive and significant in every specification, which confirms its importance for predicting product additions, at least given the set of observables we had available in our data.

With regards to the other variables in the table, urban firms seem to enjoy some benefits in terms of adding new products, though this is less obvious from the state level regressions. We were surprised to find that fixed assets were negatively correlated with adding new products, while greater amounts of "other expenses" were positively correlated. High amounts of working capital (which includes cash on hand) had the expected positive effect. Nevertheless, although these variables were statistically significant, they seem to have minimal economic significance. Each of these variables enters the regression in billions of rupees, but recall from Table 1 that the average firm only had only ₹112 million in fixed assets, ₹20 million in working capital, ₹12 million in total wages, and ₹17 million in other expenses. Thus, the typical firm would likely experience a negligible impact on product additions from each of these effects.

Larger firms, in terms of the average number of workers, also seemed to experience gains in the probability of adding products. However, this variable too would seem to have a trivial impact for the majority of firms. Average number of workers in the regression is expressed in hundreds of workers, whereas the average firm had only 160 workers in the year 2003 (the baseline year for the regressions). This means that the average firm would only experience an increase of 0.0064 ( $= 1.6 \times 0.004$ ) in the log odds of adding a product from its larger labor force. Wages are rarely significant in the regressions, ostensibly due to correlation with the size of the labor force.

**Table 4 – State-Level Logistic Regressions (RCA = 0 cutoff)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0
$\rho^0$	4.556*** (0.263)		4.029*** (0.172)	
$\omega^0$		5.814*** (0.976)		10.37*** (1.331)
urbancode	1.278** (0.486)	-0.362 (0.443)	1.071** (0.339)	-0.271 (0.342)
fixedassets	-3.971 (4.411)	-1.324 (3.526)	1.398 (2.127)	-2.880 (2.670)
workingcapital	39.35** (12.81)	6.675 (12.29)	14.74 (7.984)	23.28*** (6.339)
avg_workers	2.674*** (0.477)	-0.769 (0.507)	1.691*** (0.354)	-0.431 (0.282)
wages	-99.33 (67.89)	44.11 (33.50)	-133.2* (59.76)	84.07** (29.35)
other_exp	22.27 (38.90)	-0.916 (24.53)	44.95 (33.26)	-35.50* (15.57)
Constant	Yes	Yes	Yes	Yes
N	3836	3498	141924	86471
pseudo-R <sup>2</sup>	0.245	0.202	0.172	0.201

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2003. New products are those added over the period 2004 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 5 – State-Level Logistic Regressions (RCA = 1 cutoff)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1
$\rho^1$	3.396*** (0.514)		2.215*** (0.168)	
$\omega^1$		10.68*** (1.020)		16.12*** (2.249)
urbancode	0.647 (0.353)	-0.183 (0.342)	1.069** (0.328)	0.127 (0.321)
fixedassets	-4.142* (2.012)	2.211 (1.827)	-0.724 (1.821)	-0.612 (1.282)
workingcapital	12.35 (9.262)	2.392 (5.144)	14.81* (7.487)	22.89*** (5.855)
avg_workers	1.207*** (0.323)	-0.405 (0.228)	1.497*** (0.361)	-0.00318 (0.270)
wages	-89.34 (62.87)	53.47 (34.25)	-133.4* (61.37)	38.24 (36.37)
other_exp	48.76 (32.11)	-34.36 (24.21)	61.70 (33.20)	-35.33* (17.21)
Constant	Yes	Yes	Yes	Yes
N	6637	6299	150950	95317
pseudo-R <sup>2</sup>	0.075	0.118	0.099	0.144

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2003. New products are those added over the period 2004 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 6 – Firm-Level Logistic Regressions (RCA = 0 cutoff)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0
$\rho^1$	13.85*** (0.142)		21.30*** (0.204)	
$\omega^1$		23.97*** (0.209)		36.46*** (0.365)
urbancode	0.203*** (0.0224)	0.193*** (0.0225)	0.0601** (0.0183)	0.0572** (0.0183)
fixedassets	-0.0229** (0.00836)	-0.0229** (0.00774)	-0.0254*** (0.00766)	-0.0270*** (0.00779)
workingcapital	0.0982** (0.0353)	0.0825* (0.0342)	0.113*** (0.0313)	0.110*** (0.0269)
avg_workers	0.00328** (0.00124)	0.00353** (0.00122)	0.00488*** (0.00126)	0.00489*** (0.00126)
wages	0.145 (0.112)	0.107 (0.109)	0.101 (0.115)	0.0757 (0.120)
other_exp	0.397*** (0.115)	0.389*** (0.110)	0.409*** (0.104)	0.413*** (0.100)
Constant	Yes	Yes	Yes	Yes
N	7515654	7198032	150205855	97481419
pseudo-R <sup>2</sup>	0.080	0.094	0.085	0.096

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2003. New products are those added over the period 2004 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 7 – Firm-Level Logistic Regressions (RCA = 1 cutoff)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1
$\rho^1$	14.13*** (0.171)		21.21*** (0.199)	
$\omega^1$		24.80*** (0.228)		36.30*** (0.368)
urbancode	0.220*** (0.0237)	0.210*** (0.0238)	0.0751*** (0.0188)	0.0712*** (0.0188)
fixedassets	-0.0209** (0.00786)	-0.0208** (0.00713)	-0.0244** (0.00796)	-0.0258** (0.00819)
workingcapital	0.0970** (0.0350)	0.0872** (0.0329)	0.110*** (0.0330)	0.104*** (0.0298)
avg_workers	0.00331** (0.00125)	0.00365** (0.00123)	0.00480*** (0.00126)	0.00481*** (0.00126)
wages	0.151 (0.112)	0.119 (0.109)	0.108 (0.116)	0.0851 (0.120)
other_exp	0.379** (0.121)	0.370** (0.117)	0.404*** (0.107)	0.408*** (0.104)
Constant	Yes	Yes	Yes	Yes
N	7521509	7203887	150210253	97485812
pseudo-R <sup>2</sup>	0.076	0.092	0.085	0.096

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2003. New products are those added over the period 2004 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

The proximity measures, by comparison, have very high coefficients in the table, especially considering that both the proximity to the nearest product  $\rho$  and the average proximity  $\omega$  are constrained to be between 0 and 1. Out of the variables we had available, proximity seems to be the best predictor of new product additions, both for production and development of revealed comparative advantage. Moreover, the measure seems both statistically and economically significant.

## 5 Conclusion

This paper sought to discover whether Hidalgo et al.'s (2007) method of measuring distances between products using co-exporting correlations could explain the development of new products in India as well as it can explain the development of products at the international level. I created several potential versions of the proximity matrix for India, not based on exports of the country, but on production at the firm and state level. We then compared the predictive power of our Indian proximity matrix to the predictive power of Hidalgo et al.'s (2007) international proximity matrix, using many of the same tests they used in their paper.

The results speak strongly in favor of their method. In nearly all cases, the Indian proximity matrix proved useful for predicting which products would be added by producers over a five-year period at both the firm and the state level. The results were robust at both the 3-digit and 5-digit levels of product aggregation, and proved useful for explaining both the development of revealed comparative advantage (relative to other producers within India), as well as the simple production of products by firms. Even controlling for a number of firm characteristics, the proximity matrix calculated using the methodology of Hidalgo et al. (2007) seemed to be the best predictor of which products would be added by states and firms in India.



Despite these promising results, we acknowledge several weaknesses in our approach. While the Hidalgo measure is useful for predicting the overall pattern of product additions across the whole economy, it offers no method of identifying which firms will choose to add products. It does not take into account the profitability of potential products or the overall profitability of the firm. It also is not informative for telling us *why* certain products are co-produced. As an output-based metric, the raw proximity measure offers no insight into what might be causing these co-production correlations to emerge.

A static approach to answering the latter question is found in the second chapter of this dissertation (Flagge, 2015). In the next chapter, we will investigate some possible characteristics of firms and products that give rise to the proximity matrix in a reduced form, static setting. In the following chapter (Flagge and Chaurey, 2015), we will attempt to address the other critiques. We will employ a structural dynamic model that incorporates profits and costs, as well as several potential underlying features of the relationship between firms and products. We will then be able to model firm product addition not as a response to unobservable features linking products, but rather as a profit-maximizing decision by rational firms responding to fundamental characteristics of the relationship they have to a potential product. This method will also present a dynamic explanation for how the proximity matrix may emerge as a result of firms adding products in response to underlying characteristics.

## 6 References

Balassa, Bela (1965). "Trade Liberalisation and 'Revealed' Comparative Advantage." *The Manchester School*. Volume 33, Issue 2. 99-123.

Balassa, Bela (1977). "Revealed Comparative Advantage Revisited: An Analysis of Relative Export Shares of the Industrial Countries, 1953-1971." *The Manchester School*. Volume 45, Issue 4. 327-344.

Boschma, Ron, Asier Minondo, and Mikel Navarro (2013). "The Emergence of New Industries at the Regional Level in Spain: A Proximity Approach Based on Product Relatedness." *Economic Geography*. 89 (1) 29-51.

Colombelli, Alessandra, Jackie Krafft, and Francesco Quatraro (2014). "The emergence of new technology-based sectors in Europe: A proximity-based analysis of nanotechnology." *Research Policy*. 43. 1681-1696.

Feenstra, Robert, Robert Lipsey, Haiyan Deng, Alyson Ma, Hengyong Mo (2005). "World Trade Flow: 1962-2000." NBER Working Paper 11040.

Flagge, Matthew and Ritam Chaurey (2015). "Firm-Product Linkages and the Evolution of Product Scope." Manuscript.

Frenken, Koen and Ron Boschma (2007). "A Theoretical Framework for Evolutionary Economic Geography: Industrial Dynamics and Urban Growth as a Branching Process." *Journal of Economic Geography*. 7. 635-649.

Heckscher, Eli and Bertil Ohlin (1991). *Heckscher-Ohlin Trade Theory*. The MIT Press. Cambridge.

Hidalgo, Klinger, Barabasi, and Hausmann (2007). "The Product Space Conditions the Development of Nations." *Science*. 317, 482.

Leamer, Edward (1984). *Sources of International Comparative Advantage: Theory and Evidence*. MIT Press. Cambridge, Massachusetts.

Marshall, Alfred (1920). *Principles of Economics*. 8<sup>th</sup> Edition. London: MacMillan and Co., Ltd.

Maskell, Peter and Anders Malmberg (2007). "Myopia, knowledge development, and cluster evolution." *Journal of Economic Geography*. 7, pp. 603-618.

Neffke, Frank, Martin Henning, and Ron Boschma (2011). "How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions." *Economic Geography*. Vol. 87, Issue 3.

Neffke, Frank and Martin Henning (2008). "Revealed Relatedness: Mapping Industry Space." PEEG #08.19

Nelson, Richard and Sidney Winter (1982). *An Evolutionary Theory of Economic Change*. Cambridge, Massachusetts: Harvard University Press.

Penrose, Edith (1959). *The Theory of the Growth of the Firm*. Oxford: Basil Blackwell.

Poncet, Sandra and Felipe de Waldemar (Forthcoming). "Product Relatedness and Firm Exports in China." *World Bank Economic Review*.

Ricardo, David (1817). *On the Principles of Political Economy and Taxation*. London.

Winter, Sidney G. (2006). "Toward a neo-Schumpeterian theory of the firm." *Industrial and Corporate Change*. Vol. 15, No. 1, pp. 125-141.

## 7 Appendix

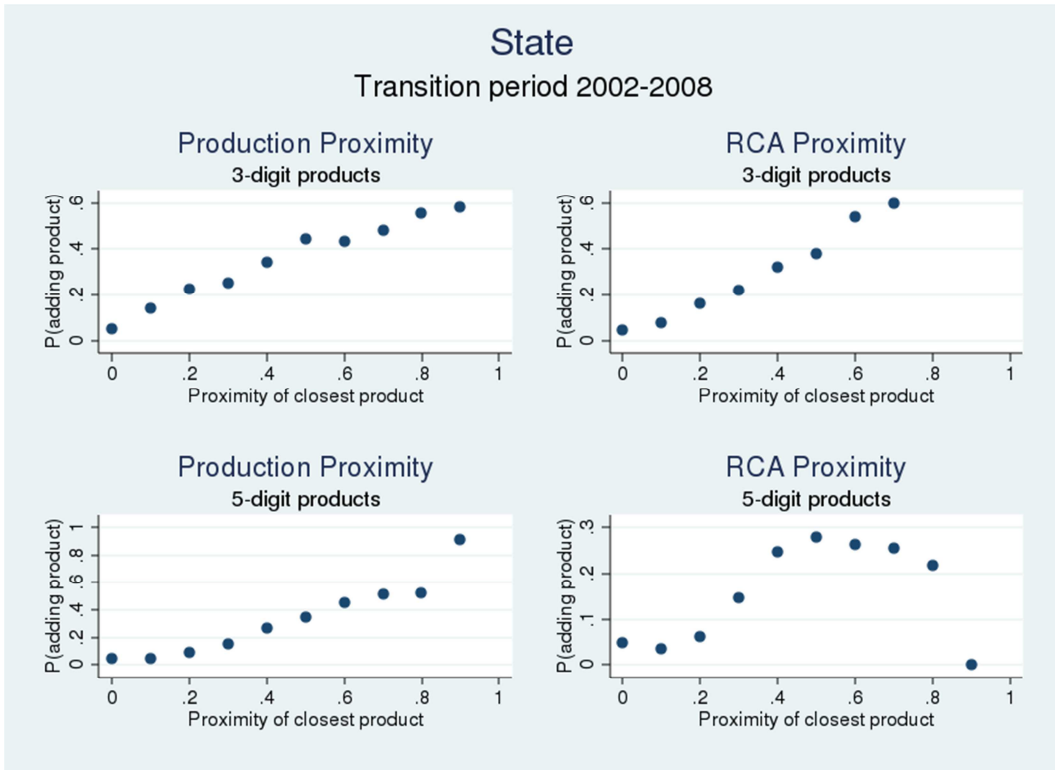
As a robustness check, we ran the same tests performed in the paper on an extended transition period covering 2002 – 2008, or two additional years. We used the same proximity matrix from the main section of the paper. The results are presented here for interested readers.

As expected, they do not differ significantly from the results in the main section of the paper. One encouraging feature is that the probability of developing RCA in a product does seem monotonically increasing for 3-digit products at the state level, as opposed to the decrease in the probability observed in the analogous graph for the baseline specification. However, there is no noticeable difference when looking at the graph for 5-digit products.

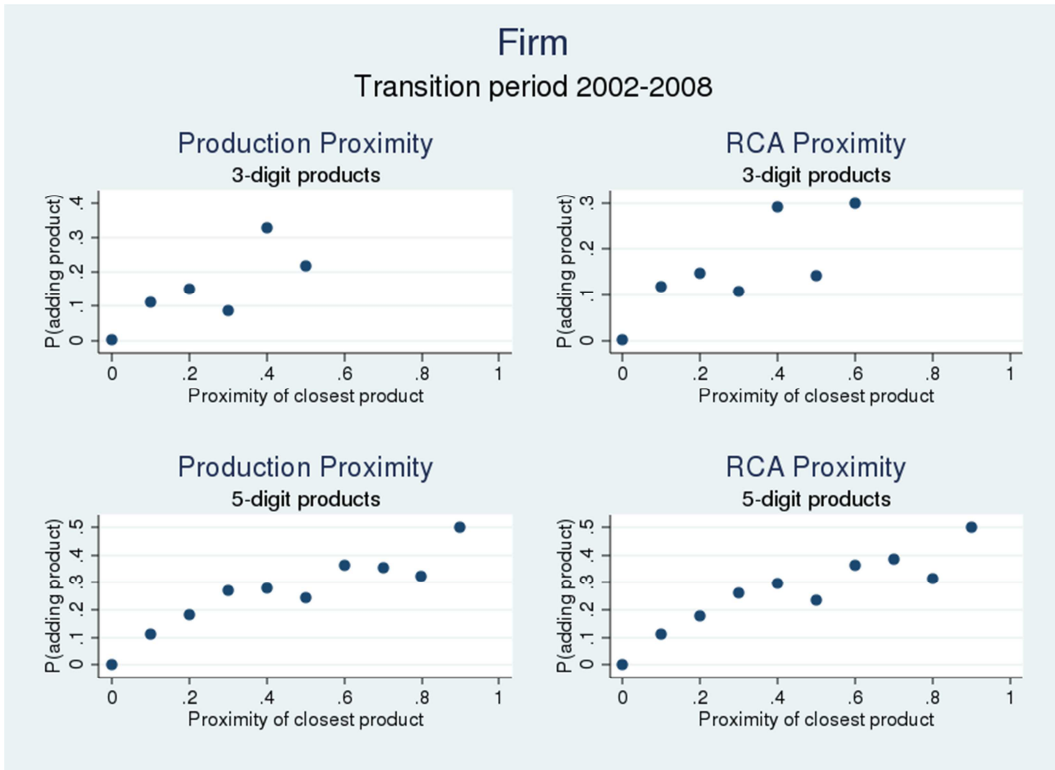
For the firm graphs in Figure 6, the 5-digit products also show a higher probability of product additions for the products at a proximity of 0.9 than they do in the baseline specification shown in Figure 2, more strongly supporting the hypothesis of a monotonically increasing probability.

The density graphs in Figure 7 and Figure 8 do not seem qualitatively different from the baselines in Figure 3 and Figure 4, neither do the logistic regressions seem to offer any insights that would contradict our summary for the baseline specification in the paper.

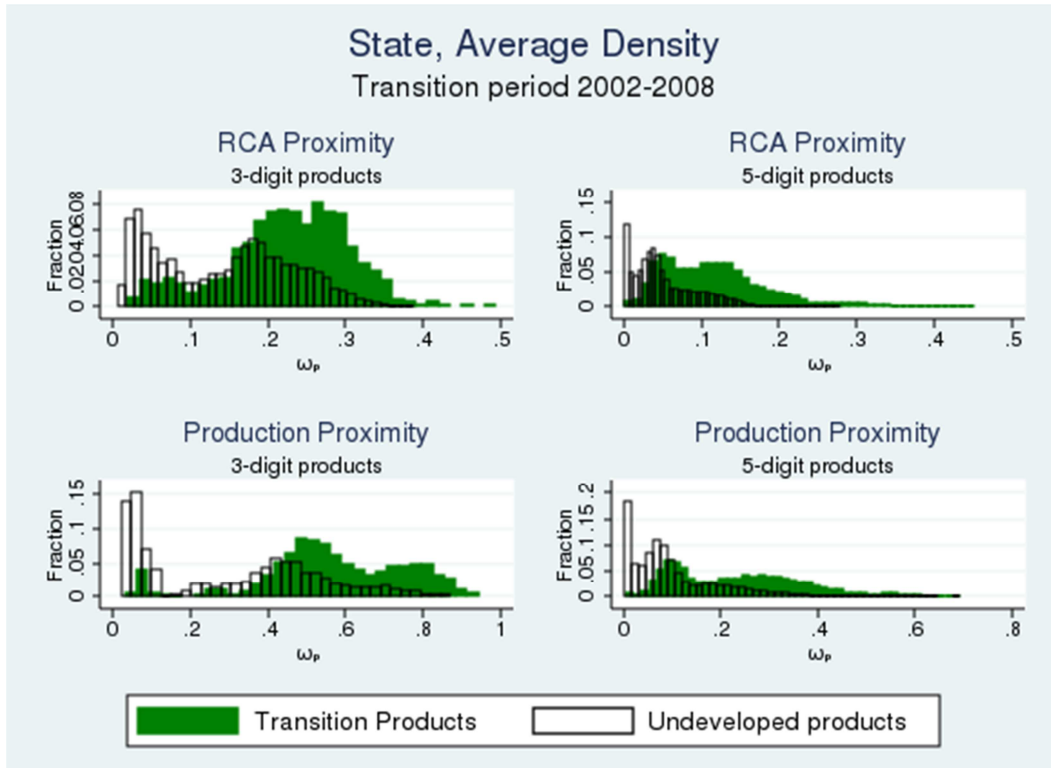
**Figure 5 - Conditional Probability of Adding Products (State)**



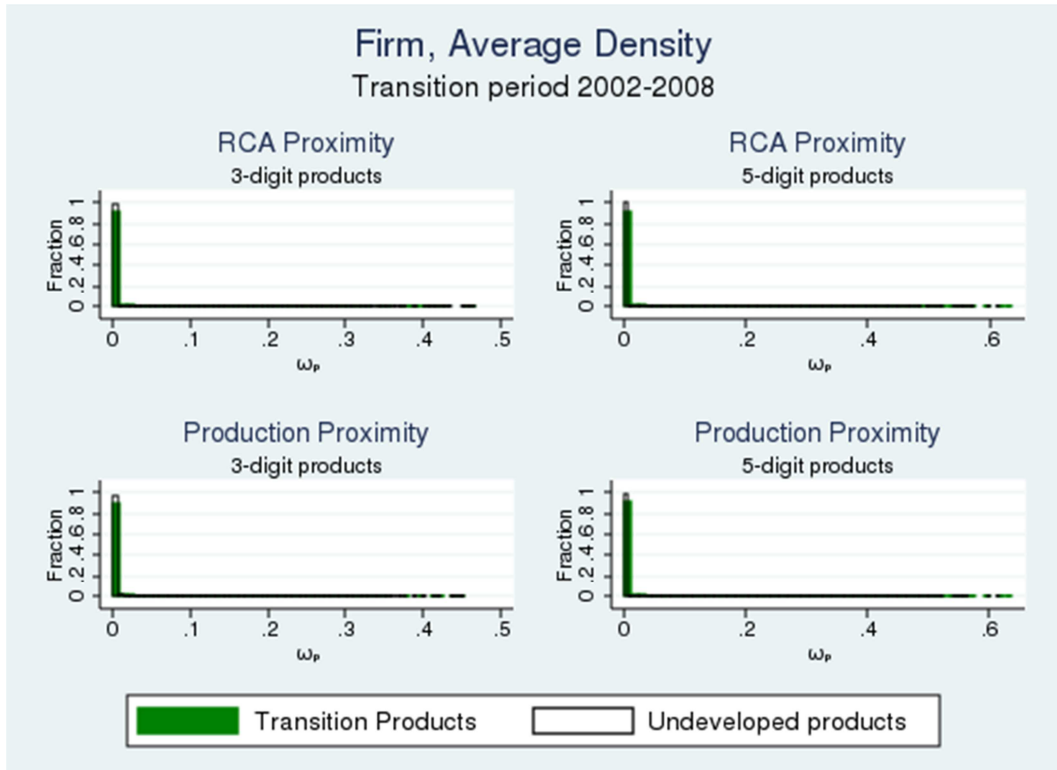
**Figure 6 - Conditional Probability of Adding Products (Firm)**



**Figure 7 - Density of Productive Structure around Transition vs. Undeveloped Product (State)**



**Figure 8 - Density of Productive Structure around Transition vs. Undeveloped Product (Firm)**





**Table 8 – State-level logistic regressions (RCA cutoff = 0)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0
$\rho^0$	4.603*** (0.289)		4.414*** (0.157)	
$\omega^0$		7.163*** (0.996)		12.15*** (1.305)
urbancode	1.943** (0.708)	-0.823* (0.375)	1.224*** (0.358)	-0.528 (0.419)
fixedassets	-9.823* (4.976)	6.047 (4.539)	-2.585 (3.628)	0.392 (4.326)
workingcapital	56.45* (24.67)	11.82 (15.57)	5.363 (17.52)	34.36** (13.05)
avg_workers	2.002*** (0.547)	-0.951* (0.443)	1.473*** (0.350)	-0.383 (0.262)
wages	-178.8* (91.00)	96.26 (65.69)	-182.2* (73.16)	66.40** (23.50)
other_exp	119.5 (61.22)	-114.5* (49.93)	128.2* (52.74)	-83.83** (32.30)
Constant	Yes	Yes	Yes	Yes
N	3972	3637	142949	87286
pseudo-R <sup>2</sup>	0.247	0.243	0.187	0.226

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2001. New products are those added over the period 2002 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 9 – State-level logistic regressions (RCA cutoff = 1)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1
$\rho^0$	3.938*** (0.530)		2.697*** (0.223)	
$\omega^0$		14.28*** (1.255)		26.48*** (2.291)
urbancode	0.846* (0.366)	-0.834* (0.365)	1.138*** (0.328)	-0.608 (0.453)
fixedassets	-6.341* (3.213)	3.465 (3.266)	-3.103 (3.124)	1.180 (3.815)
workingcapital	19.54 (14.64)	-1.931 (10.99)	4.318 (15.73)	37.65* (14.91)
avg_workers	0.929*** (0.264)	-0.436 (0.226)	1.248*** (0.305)	-0.286 (0.204)
wages	-93.16 (48.18)	95.93** (36.12)	-170.7** (61.64)	74.32** (23.09)
other_exp	66.29 (35.22)	-94.52*** (26.07)	129.4** (44.00)	-121.9** (43.03)
Constant	Yes	Yes	Yes	Yes
N	6696	6360	151498	95706
pseudo-R <sup>2</sup>	0.082	0.151	0.113	0.214

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2001. New products are those added over the period 2002 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 10 – Firm-level logistic regressions (RCA cutoff = 0)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0	Logit new RCA > 0
$\rho^0$	14.01*** (0.155)		22.97*** (0.245)	
$\omega^0$		25.49*** (0.249)		41.65*** (0.372)
urbancode	0.216*** (0.0232)	0.199*** (0.0232)	0.0616** (0.0197)	0.0683*** (0.0197)
fixedassets	-0.0220 (0.0128)	-0.0206 (0.0111)	-0.0220 (0.0121)	-0.0243* (0.0115)
workingcapital	0.128*** (0.0305)	0.104*** (0.0290)	0.138*** (0.0300)	0.127*** (0.0303)
avg_workers	0.00310* (0.00125)	0.00327** (0.00122)	0.00459*** (0.00124)	0.00443*** (0.00123)
wages	0.0489 (0.130)	0.0132 (0.119)	0.0186 (0.127)	0.00677 (0.132)
other_exp	0.383*** (0.0937)	0.384*** (0.0846)	0.357*** (0.0973)	0.372*** (0.0964)
Constant	Yes	Yes	Yes	Yes
N	6824447	6536022	136399902	88521386
pseudo-R <sup>2</sup>	0.082	0.100	0.093	0.113

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2001. New products are those added over the period 2002 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

**Table 11 – Firm-level logistic regressions (RCA cutoff = 1)**

Regressor	3-digit products		5-digit products	
	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1	Logit new RCA > 1
$\rho^0$	14.13*** (0.185)		22.91*** (0.243)	
$\omega^0$		25.79*** (0.275)		41.71*** (0.375)
urbancode	0.257*** (0.0245)	0.242*** (0.0246)	0.0846*** (0.0201)	0.0906*** (0.0202)
fixedassets	-0.0240 (0.0131)	-0.0233* (0.0117)	-0.0226 (0.0124)	-0.0242* (0.0118)
workingcapital	0.130*** (0.0326)	0.113*** (0.0302)	0.138*** (0.0311)	0.123*** (0.0324)
avg_workers	0.00286* (0.00127)	0.00309* (0.00125)	0.00446*** (0.00124)	0.00427*** (0.00123)
wages	0.0774 (0.129)	0.0506 (0.120)	0.0289 (0.126)	0.0221 (0.128)
other_exp	0.405*** (0.0939)	0.411*** (0.0878)	0.365*** (0.0982)	0.386*** (0.0953)
Constant	Yes	Yes	Yes	Yes
N	6830105	6541679	136404297	88525777
pseudo-R <sup>2</sup>	0.079	0.096	0.093	0.112

Notes: Standard errors clustered at the state-level in parentheses. Average number of workers expressed in hundreds of workers. Fixed assets, working capital, wages, and other expenses expressed in billions of rupees. Values of all control variables correspond to the values for that producer for the year 2001. New products are those added over the period 2002 – 2008.

\*\*\* = significant at 0.1% level. \*\* = significant at 1% level. \* = significant at 5% level.

# Chapter 2: Foundations of Product Proximity\*

Matthew Flagge<sup>1</sup>

<sup>1</sup>Columbia University

May 2015

This paper seek to uncover the characteristics of firms and products that give rise to product proximity in India as calculated by the methodology in Hidalgo et al. (2007). Proximity measures how frequently certain pairs of products are co-produced within states or firms. I examine demand complementarities, input similarity, urban environment, firm size, and labor intensity within a reduced form tobit model. We find demand complementarities to dominate the specification, with the other variables having smaller but generally positive effects on the probability that two products are co-produced.

---

\* I would like to thank Eric Verhoogen and Amit Khandelwal for helpful comments and for facilitating access to the Annual Survey of Industries data. The author accepts responsibility for any errors.

# 1 Introduction

Why are some pairs of products more likely to be produced together than others? Leamer (1984) observed product clusters in the export data and found them to be inconsistent with the standard explanations of capital or labor abundance. Bernard, Redding, and Schott (2010) observed similar clusters at the production level in the US Census of Manufactures and found that such production patterns could not be explained by a random model. The fact that the pattern is not random implies that it is being driven by *something*—some characteristic of the firms that make them especially well-adapted or inclined to producing certain pairs of products, or features of the products themselves that make them likely to be produced together.

One way of representing these clusters was provided by Hidalgo et al. (2007). Looking at export data, they calculated the probability of exporting each product, conditional on exporting every other product. The result was a matrix of these conditional probabilities which they called the “Product Space.” They found that these relationships conditioned the development path of nations. Specifically, countries were more likely to develop comparative advantage in products that were close to their existing products in the product space, as opposed to those which were distant.<sup>10</sup> Their idea became very influential. As of February 2015, the *Social Science Citation Index* placed Hidalgo et al. (2007) in the top 1% of the most cited papers in the field of Economics and Business.

Hidalgo et al. were admittedly agnostic about what might be causing these correlations. As they argue, products could potentially have many different inputs that could be required to produce them successfully. These might include labor, physical and human capital, material inputs, institutions, environmental conditions (such as a moist or dry climate), and technological or physical infrastructure,

---

<sup>10</sup> Put another way, if products A and B are typically exported together, and C is rarely exported with A, then a country that has comparative advantage in only A is more likely to develop comparative advantage in B than in C.

among many other possibilities. Rather than trying to build a model that incorporates all of these diverse features (many of which are difficult to measure or quantify), they instead assume that if two products use similar inputs (viewed in the broad sense), then a country which produces one product successfully will be well-positioned to begin producing the other product.

Hidalgo et al.'s study looked at world trade data for countries. Getting detailed data about the productive environment in which products are produced is difficult in such an aggregated and diverse setting. However, since it has been shown<sup>11</sup> that their methodology is informative for predicting product additions by firms as well as countries, we feel it would be fruitful to study the underlying characteristics of firms or products that lead to the emergence of product proximity at the firm level where more detailed and disaggregated data are available. The present study seeks to uncover how much of the proximity measure can be explained by observed factors in a rich plant-level dataset in India.

We use the Annual Survey of Industries (ASI) in India to compute a matrix of pairwise proximities measuring the likelihood that two products are produced together by the same state or firm in India. The ASI is a plant-level panel dataset containing a large number of firm characteristics. We use these to compute product characteristics such as labor intensity, material input requirements, etc., and regress the proximity measure between product pairs on these properties to see which characteristics are most correlated with the tendency of products to be produced together.

We find that our measures of demand complementarities and input similarity both seem to be most highly correlated with output co-production. These dominate in each one of our specifications. Other factors that readers would expect to be important, such as similar labor intensities and similar firm size, also have a positive and significant impact on which products are produced together.

---

<sup>11</sup> See Poncet and de Waldemar (forthcoming), as well as the other two chapters of this dissertation (Flagge 2015, and Flagge and Chaurey 2015).

The paper is organized as follows. Section 2 describes the data. Section 3 goes through how we compute each of the variables included in our analysis. Section 4 presents the results, and section 5 concludes.

## 2 Data

The primary dataset we use for this project is the *Annual Survey of Industries* (ASI) from India. This is a factory-level panel spanning 1999 – 2008, and is a representative sample of all the factories in India.<sup>12</sup> The panel includes data such as number of workers in a plant, total wages paid to each type of worker, value of plant and machinery, working capital, material inputs, outputs, and state of the plant. We merge the panel with an associated cross-section to narrow the location of the plant down to the district level.

Both inputs and outputs in the data are denoted as 5-digit product codes in the ASICC classification. When we speak of products in this paper, we are referring to products as classified by one of these 5-digit codes. This is a fairly disaggregated system of classification. There are over 5395 different products in the data. Of these, 204 occur only as outputs, 193 occur only as inputs, and the remaining 4998 occur as both inputs and outputs.

Each sector has a different number of products associated with it. Across sectors there will naturally be different levels of heterogeneity even within 5-digit codes, but we feel the products as defined are specific enough as to allow us to make inferences about their production fundamentals. Examples of 5-digit products in each industry, as well as the number of 5-digit product codes in each industry classification, can be found in Table 12.

---

<sup>12</sup> The sampling procedure was pseudo-random in that firm characteristics were taken into consideration in determining the probability of being sampled. For instance, factories in isolated areas, and those with a large number of employees were sampled with certainty, while others were sampled with only a given probability.



Another characteristic to note about our data is that over 94% of factories belong to single-factory firms. This is valuable since it means we can safely use factory characteristics to explain co-production at the factory level without having to worry about unobserved features of the firm influencing their decisions for the majority of our dataset. Because of this, we will use the terms *factory* and *firm* interchangeably in this paper.

**Table 12**

<b>Industry</b>	<b>Number of Products</b>	<b>Example of 5-digit product (code)</b>
Animal, Vegetable, Forestry	613	Cakes, pastries, & muffins (13403)
Ores, minerals, fuels, electricity	267	Diamond powder (22701)
Chemicals	1179	Raspberry Oil (36133)
Rubber, leather, plastic	473	Leather belts (44908)
Wood, cork, paper	274	Lottery tickets (56125)
Textiles	414	Cotton track suits (63435)
Metals and Machinery	1565	Electric toasters (77707)
Transportation Equipment	150	Racing cars (82141)
Miscellaneous	460	Fountain pens (95205)

### 3 Product-to-Product Relationships

In this section, we will go through how we compute each of the variables included in our regressions. Since each observation in the regression is a pair of products, each one of these variables is meant to capture some aspect of the relationship between those products.

The general regression specification is given in the following equation:

$$\phi(p, q) = \alpha + \beta^\theta \theta(p, q) + \beta^\sigma \sigma(p, q) + \beta^L \rho_L(p, q) + \beta^u \rho_u(p, q) + \beta^s \rho_s(p, q) + \epsilon \quad (6)$$

Where  $\phi$  is the proximity measure,  $\theta$  is a measure of demand complementarities,  $\sigma$  is a measure of input similarity,  $\rho_L$  is a measure of labor intensity,  $\rho_u$  is urban intensity, and  $\rho_s$  is a measure of firm size. Since each observation within this regression is a product pair, each measure is a function

of the products  $p$  and  $q$ . In the following sections we give the specific equations laying out how each of the measures is defined.

### 3.1 Output Proximity

The first variable we look at is proximity of outputs. We compute this from the formula by Hidalgo et al. (2007), who derived it to capture the idea that the ability of a producer to make one product is related to its ability to make other products. They defined proximity in terms of revealed comparative advantage (RCA), a concept invented by Balassa (1965) to capture whether or not a country exports more or less of a particular product than average. It is a normalized measure, so being a large or a small exporter should not affect it. Rather, it is based on whether a particular product represents a large or small share of total exports, relative to that product's share of total exports in the general population.

Balassa's motivation was that since the fundamentals underlying comparative advantage are difficult to observe (perhaps even difficult to know), the comparative advantage of a producer might be *revealed* by observing the patterns of trade. Using his formula, the RCA of an exporter  $e$  for a particular product  $p$  is given by:

$$RCA(e, p) = \frac{\frac{x_{ep}}{\sum_p x_{ep}}}{\frac{\sum_e x_{ep}}{\sum_{ep} x_{ep}}} \quad (7)$$

Where  $x_{ep}$  is the total exports of product  $p$  by exporter  $e$ . Based on Balassa's measure, an exporter that has  $RCA > 1$  for a particular product, exports relatively more of that product than average, and thus is said to have revealed comparative advantage in that product.

In our data, we do not observe exports, but we do observe production at the firm level. Therefore, our RCA calculations represent whether or not a firm produces more of a particular product relative to the other firms in India during that particular year.

Once we have Balassa's RCA measure, we can use it to compute the proximity between products. In Hidalgo et al. (2007), they base their proximity measure on the conditional probability of having revealed comparative advantage ( $RCA > 1$ ) in a pair of products. In order to generate a symmetric proximity measure, they take the minimum of these conditional probabilities for each product pair. Thus, the proximity  $\phi^1$  between products  $p$  and  $q$  would be:

$$\phi^1(p, q) = \min\{P(RCA(p) > 1 | RCA(q) > 1), P(RCA(q) > 1 | RCA(p) > 1)\} \quad (8)$$

We also go one step further than Hidalgo et al. and compute another proximity matrix based solely on co-production correlations. So rather than measuring how frequently firms have a revealed comparative advantage in a particular product, our second proximity matrix measures how frequently certain pairs of products are produced together *at all*. Since our analysis takes place within the bounds of a country, and studies firms that each produce only a small subset of the total universe of products, it makes sense to also consider this alternative proximity matrix.

We will denote this second type of proximity as  $\phi^0$ , where the superscript 0 represents the threshold of RCA needed in proximity formula. Thus, the proximity between products in our alternative proximity matrix is given by:

$$\phi^0(p, q) = \min\{P(RCA(p) > 0 | RCA(q) > 0), P(RCA(q) > 0 | RCA(p) > 0)\} \quad (9)$$

Equation (9) is analogous to equation (8), except for the RCA cutoff.

The proximity matrices are computed separately for each year. Then, to reduce measurement error, Hidalgo et al. generated their baseline proximity matrix by averaging the matrices for three different years. In their data, they created their preferred matrix by averaging the years 1998 – 2000. We use the same procedure to create our preferred matrix, which we generate by averaging the year-specific matrices for the years 2001 – 2003.

### **3.2 Input Proximity (Demand Complementarities)**

One potential explanation for why products certain products might be produced together by the same firm is that the firm is catering to demand. Bernard, Redding, and Schott (2010) suggested that demand-side complementarities might help explain the non-random patterns of co-production in U.S. data. We do not have data on consumer demand that can be easily linked to our current dataset. However, what we can observe is one portion of demand: the demand for products as intermediate inputs.

One might imagine that if producers demand certain products in pairs, then suppliers of those products further up the supply chain might want to produce both products in order to sell more of their output to the same consumers, where in this case, “consumers” are merely producers further down the supply chain.

To capture this notion, we need a measure of how often certain pairs of products are bought together by the same firms. Fortunately, we already have a formula to express this—the proximity. For this measure of product relatedness, we therefore use exactly the formulas specified in section 3.1, but for inputs instead of outputs. We will call this quantity either  $\theta^1$  or  $\theta^0$ , again using the superscripts to denote the cutoff for the RCA measure in equation (9).

### 3.3 Input Similarity (Cost Complementarities)

Another potential explanation for why certain products may be produced together is cost complementarities in terms of inputs. We might imagine that there are costs associated with using certain inputs in addition to the raw material cost. For instance, a firm might incur a search cost to finding a supplier, or a learning cost to discover how to use a certain type of input effectively. If we were to model this (which we won't) we would say that each input has a fixed cost as well as a marginal cost. In such a world, firms could achieve economies of scale by producing products that share inputs, since the firm would only have to pay the fixed input cost once for each type of input it uses.<sup>13</sup>

To measure the degree to which two products share inputs, we use the product dissimilarity component of Gollop and Monahan's (1991) index of diversification. This captures the degree to which the inputs of two products are similar in terms of the cost shares of those inputs. For any two products,  $p$  and  $q$ , the Gollop and Monahan index of product dissimilarity is given by:

$$\sigma'(p, q) = \left( \sum_j \frac{|w_{pj} - w_{qj}|}{2} \right)^2 \quad (10)$$

Where  $w_{pj}$  is the cost share of the  $j^{th}$  input in the  $k^{th}$  product. This produces an index between 0 and 1 such that  $\sigma'(p, q) = 1$  means that  $p$  and  $q$  share no inputs and  $\sigma'(p, q) = 0$  means they are identical in terms of input cost shares. We invert this measure by defining:

$$\sigma(p, q) = |\sigma'(p, q) - 1| \quad (11)$$

So higher values of  $\sigma_{pq}$  correspond to more similar products. This adjustment is merely for ease of interpretation, so that positive effects of input similarity on proximity will have positive coefficients in the regression tables.

---

<sup>13</sup> This idea is similar to the intuition laid out in Penrose (1959).

Determining the input cost shares associated with each product requires a product-level input-output table. Typically such tables are only available at the industry level, which is too coarse for our particular application. However, since we have data on inputs and outputs for each firm, we compute our own input-output table using single-product firms, which allows us to assign inputs to output unambiguously.<sup>14</sup> This gives us a ratio of the number of inputs of each type needed per unit of output for individual firms. The recipe for that product in the input-output table is then the mean of the firm-specific recipes for each firm that produces that product (and only that product) across all the years of the sample.

### **3.4 Labor Intensity**

Classical trade theories (such as Heckscher and Ohlin [1991]) suggest that what drives product choice could be relative factor intensities across country borders. That is, capital intensive countries should export capital intensive goods (and vice versa). This would suggest that two products with similar labor intensities might be produced by the same producer. The intuition and mechanism for this is not fundamentally different from what could be driving product co-production due to similarity of intermediate inputs, as discussed in the previous section. However, to add more depth into the exact

---

<sup>14</sup> Another alternative found in the literature (such as Foster, Haltiwanger, and Syverson [2008] and DeLoecker [2011] among others) is to assign inputs to outputs proportionally to each output's share of firm revenues. This would allow us to include multi-product firms in our construction of the input-output table. Although this has desirable properties, for this particular setting it would be inappropriate because that methodology introduces a source of spurious correlation with output proximity into the input similarity measure. Since we are using output proximity as our dependent variable, employing that method of building an input-output table would therefore be unacceptable here. The reason is that if two products are produced in the same firm, then according to this method, they will necessarily be assigned the same inputs, in the same proportions, at that firm. This will make it appear like they have almost identical inputs for that firm-year. Aggregating across all firm-years to get the product input profile will not necessarily remove this source of endogeneity. We nevertheless performed this exercise anyways, constructed such a measure, and ran our regressions with the multi-product firm version of the input-output table. The results of this fool's errand, not reported in the paper, confirmed our suspicions about the endogeneity. The input similarity measure in those regressions dominated every other variable by a large amount, and the coefficients were substantially larger than any we observed in the comparable regressions in the paper.

nature of the input similarity, and to pay homage to neoclassical economics which divides inputs into capital and labor, we calculate a labor intensity for each product in our dataset.

To do this, we first find the total amount spent on wages and capital for each firm-year.<sup>15</sup> As we did for intermediate inputs, we then allocate these expenses proportionally to each of the products the firm produces, based on the product's share of output. This allows us to compute a (firm-year)-specific labor intensity for each product the firm produces. We then take the median across all firms and years to determine the product-specific labor intensities.

However, what we need is to measure the distance between these labor intensities, and identify products that have very similar labor intensities. We therefore map the difference between the product labor intensities onto the (0,1] space using the exponential function, and we define the labor intensity proximity between two products as<sup>16</sup>:

$$\rho_L(p, q) = e^{-|l_p - l_q|} \quad (12)$$

Where  $l_p$  is the labor intensity of product  $p$  and  $||$  represents the absolute value. This function takes a value of 1 if two products have identical labor intensities, and asymptotically approaches 0 as they become more dissimilar.

---

<sup>15</sup> What we define as wages excludes bonuses, contributions to the Provident Fund (social security), and welfare. Capital is defined as the closing value of plant and machinery in the given year.

<sup>16</sup> We chose the exponential as opposed to a linear mapping function because of our prior about how similar two products' labor intensities would have to be in order to *actually* be similar products. Labor intensity is a ratio, so small differences in labor intensity can correspond to large differences in terms of needed inputs. Therefore, we sought a function that dropped in value quickly with increases in  $|l_p - l_q|$ . Other potential scaling functions, such as  $\frac{1}{|l_p - l_q| + 1}$  or  $1 - \frac{2 \text{Arctan}|l_p - l_q|}{\pi}$  behave similarly to our chosen function, but the exponential was selected because of its simplicity.

### 3.5 Urban Share

Some authors<sup>17</sup> have suggested that certain products may be better produced within the diversity of a city as opposed to a rural environment. We incorporate this notion into our estimation by looking at the proportion of each product that is produced by a factory in an urban setting.

We weight this in two ways: by factories, and by sale value. The factory-weighted urban share counts each firm-year in which a product is produced as one observation, and calculates the proportion of these in which the product was produced in an urban setting. The value-weighted urban share calculates the proportion of the total sale value of a product that is produced in an urban setting.

The proximity between two products along this dimension is then given by:

$$\rho_u(p, q) = 1 - |u_p - u_q| \quad (13)$$

Where  $u_p$  is the urban share of product  $p$  with either factory or value-weighted measure. This will yield a measure in which products with the same urban share have a proximity of 1, and products with opposite urban shares will have a proximity of 0.

### 3.6 Firm Size

Other authors<sup>18</sup> have suggested that more innovative products might be more likely to be produced by firms of a certain size, as measured either by number of employees or total value of their output. We therefore calculate the number of employees and total sales of the “typical” firm that produces each product, and incorporate that into our estimation.

---

<sup>17</sup> Jacobs (1969) first put forth this hypothesis. Duranton and Puga (2001) offered a theoretical model explaining why more innovative products might be better produced in a diversified urban environment. On the empirical side, Feldman and Audretsch (1999) have found that areas with a greater diversity of economic activities promote innovation.

<sup>18</sup> Scherer (1965) found that inventive output, as measured by patents, was increasing in firm sales. Ettlie and Rubenstein (1987) found that larger firms, as measured by the number of employees, are better able to commercialize radical new products successfully, but that this effect is diminished for very large firms.



For each firm-product, we compute the total sale value in real terms of all the firm's products in that year, as well as the average number of workers employed by that firm, excluding the owner and non-paid employees. We then take the average of these across all years and firms to get the values we assign to the products. To map these numbers into product proximities, we use a linear mapping function:

$$\rho_s(p, q) = 1 - \frac{|s_p - s_q|}{\bar{s}_d - \underline{s}_d} \quad (14)$$

Where  $s_p$  is the firm size (measured either in revenue or workforce) associated with product  $p$ , and  $\bar{s}_d$  and  $\underline{s}_d$  are the maximum and minimum values respectively of  $|s_p - s_q|$  found in the data. This will map the differences in firm size into  $[0,1]$  space, with products coming from firms of the same size having a proximity of 1, and products coming from firms of radically different sizes having proximity of 0 (or close to it).

## 4 Results

As a first pass, we report the results of OLS regressions of proximity on the explanatory variables discussed in the previous section. The results for the state-level proximity matrix are given in Table 13 and Table 14. The results for the firm-level proximity matrix are given in Table 15 and Table 16. The dependent variable in each of the regressions is the  $ij^{th}$  element of the given proximity matrix, either  $\phi^0$  or  $\phi^1$ . Each observation is linked to an element in the lower triangle of that matrix (excluding the diagonal), so each observation represents a product pair.<sup>19</sup>

In each table, the demand complementarities represented by the input proximity matrix (either  $\theta^0$  or  $\theta^1$ ) seem to be most highly correlated with product proximity. By this we mean that the

---

<sup>19</sup> Because we are dealing with dyadic data, all of the standard errors in the tables in this paper are calculated using the multi-dimensional clustering formula found in Cameron, Gelbach, and Miller (2011). This method was also used by Ana Fieler (2011) to handle trade data, which is also dyadic.

coefficients on the demand complementarities are the highest, and still very significant. For the firm-level regression, it seems to be the *only* meaningful variable in the regression. Most of the other variables, while statistically different from zero, are dwarfed in magnitude by the demand complementarities.<sup>20</sup> The lone exception is the input similarity measure.

The other regressors feature more prominently in the state-level regression. It should be noted that the state-level proximity matrix has more non-zero elements in it, meaning there is more room for the explanatory variables to get some leverage in the regression.

Again, the demand complementarities seem to dominate the table. These are followed in importance by products produced in the same urban (or rural) setting, and input similarity across products. Somewhat surprisingly, the unweighted urban measure seemed to have more impact than the urban measure weighted according to sale value, denoted *weightedurban* in the tables. Similarity of labor intensity, denoted by *laborint*, displayed a small but positive impact in explaining the proximity matrix based on revealed comparative advantage ( $\phi^1$ ), but a negligible impact for explaining the matrix for production ( $\phi^0$ ).

Firm size, as measured by number of workers (denoted *numworkers*) in the table, was positively correlated with proximity. However, we were surprised to find that firm size as measured by revenues (denoted *firmrev*) was *negatively* correlated, albeit often in a manner that is statistically insignificant. We suspected that perhaps this was due to revenues being correlated with number of workers, however the negative correlation with proximity persists even in the regressions in which *numworkers* is excluded.

---

<sup>20</sup> We consider it relevant to talk about the magnitudes of the coefficients here since all of our regressors have been mapped into the [0,1] space.

**Table 13 – State-level OLS Regressions (RCA cutoff = 1)**

Regressor	(1) $\phi^1$	(2) $\phi^1$	(3) $\phi^1$	(4) $\phi^1$	(5) $\phi^1$
$\theta^1$	152.229*** (6.0663)	152.892*** (6.0775)	112.094*** (3.3096)	112.491*** (3.3130)	152.251*** (6.0603)
inputs	67.625*** (6.8526)	68.227*** (6.8666)			67.773*** (6.8452)
laborint	6.872** (2.1893)	7.115** (2.1909)	7.233*** (1.1630)	7.192*** (1.1639)	6.951** (2.1877)
urban	71.672*** (3.8921)	71.494*** (3.8947)	54.115*** (1.9399)	54.140*** (1.9403)	61.516*** (5.0269)
numworkers	73.410*** (21.0230)		30.309*** (7.8044)		72.337*** (20.9620)
firmrev	-33.144 (36.3304)	-1.042 (31.1474)	-43.746*** (8.7285)	-32.920*** (7.7449)	-32.725 (36.2341)
weightedurban					12.558** (3.9018)
constant	-0.025 (0.0335)	0.015 (0.0311)	0.023** (0.0089)	0.042*** (0.0078)	-0.026 (0.0334)
N	2045199	2045199	11600937	11600937	2045199
R-sq	0.048	0.047	0.029	0.029	0.048

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

**Table 14 – State-Level OLS Regressions (RCA cutoff = 0)**

Regressor	(1) $\phi^0$	(2) $\phi^0$	(3) $\phi^0$	(4) $\phi^0$	(5) $\phi^0$
$\theta^0$	320.964*** (10.8809)	321.257*** (10.8813)	262.971*** (6.1769)	263.433*** (6.1784)	320.758*** (10.8600)
inputs	106.151*** (11.5041)	106.664*** (11.5165)			106.400*** (11.4935)
laborint	-0.333 (5.0817)	-0.144 (5.0760)	-0.591 (2.2092)	-0.670 (2.2104)	-0.172 (5.0730)
urban	145.258*** (9.0845)	145.100*** (9.0858)	96.098*** (3.5430)	96.107*** (3.5435)	127.864*** (12.1273)
numworkers	60.233 (42.3425)		41.424*** (12.5326)		58.421 (42.2271)
firmrev	-39.470 (52.9296)	-13.101 (49.9223)	-57.910*** (14.5966)	-43.093** (13.2649)	-38.786 (52.6774)
weightedurban					21.548* (9.1117)
constant	-0.013 (0.0583)	0.020 (0.0502)	0.014 (0.0149)	0.040** (0.0135)	-0.014 (0.0581)
N	2045199	2045199	11600937	11600937	2045199
R-sq	0.189	0.189	0.145	0.145	0.190

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

**Table 15 – Firm-Level OLS Regressions (RCA cutoff = 0)**

Regressor	(1) $\phi^0$	(2) $\phi^0$	(3) $\phi^0$	(4) $\phi^0$	(5) $\phi^0$
$\theta^0$	170.533*** (27.7424)	170.541*** (27.7432)	139.364*** (10.5464)	139.393*** (10.5468)	170.531*** (27.7430)
inputs	10.222*** (1.0683)	10.227*** (1.0683)			10.222*** (1.0684)
laborint	0.134*** (0.0171)	0.136*** (0.0171)	0.181*** (0.0104)	0.180*** (0.0104)	0.134*** (0.0171)
urban	0.331*** (0.0265)	0.330*** (0.0265)	0.296*** (0.0110)	0.296*** (0.0110)	0.350*** (0.0342)
numworkers	0.538*** (0.1273)		0.385*** (0.0524)		0.540*** (0.1274)
firmrev	-0.125 (0.0992)	0.110 (0.1091)	-0.233*** (0.0671)	-0.096 (0.0595)	-0.126 (0.0991)
weightedurban					-0.023 (0.0230)
constant	-0.001*** (0.0001)	-0.000** (0.0001)	-0.000*** (0.0001)	-0.000 (0.0001)	-0.001*** (0.0001)
N	2045199	2045199	11600937	11600937	2045199
R-sq	0.074	0.074	0.024	0.024	0.074

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

**Table 16 – Firm-Level OLS Regressions (RCA cutoff = 1)**

Regressor	(1) $\phi^1$	(2) $\phi^1$	(3) $\phi^1$	(4) $\phi^1$	(5) $\phi^1$
$\theta^1$	152.229*** (6.0663)	152.892*** (6.0775)	112.094*** (3.3096)	112.491*** (3.3130)	152.251*** (6.0603)
inputs	67.625*** (6.8526)	68.227*** (6.8666)			67.773*** (6.8452)
laborint	6.872** (2.1893)	7.115** (2.1909)	7.233*** (1.1630)	7.192*** (1.1639)	6.951** (2.1877)
urban	71.672*** (3.8921)	71.494*** (3.8947)	54.115*** (1.9399)	54.140*** (1.9403)	61.516*** (5.0269)
numworkers	73.410*** (21.0230)		30.309*** (7.8044)		72.337*** (20.9620)
firmrev	-33.144 (36.3304)	-1.042 (31.1474)	-43.746*** (8.7285)	-32.920*** (7.7449)	-32.725 (36.2341)
weightedurban					12.558** (3.9018)
constant	-0.025 (0.0335)	0.015 (0.0311)	0.023** (0.0089)	0.042*** (0.0078)	-0.026 (0.0334)
N	2045199	2045199	11600937	11600937	2045199
R-sq	0.048	0.047	0.029	0.029	0.048

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

Nevertheless, we are hesitant to accept the results from the OLS regression at face value due to the nature of our dependent variable. Proximity is a probability, and is therefore constrained to be between 0 and 1. A linear regression model is therefore likely to make poor predictions, especially around the bounds. Traditional models for dealing with explanatory variables between 0 and 1, such as logit and probit, are not well suited to situations when the dependent variable can take on more than two values. We therefore use a Tobit model.

The state-level regressions are found in Table 17, and the firm-level regressions are found in Table 18 and Table 19.<sup>21</sup> The Tobit model seems to explain the data better than OLS, judging by the higher R-squared we see in the tables. With the Tobit model, the firm-level regressions seem to perform much better. This implies that our regressors are better predictors of product proximity at the firm level. This is what one might expect if one believes that the foundations of proximity can be found largely within the boundaries of the firm.

---

<sup>21</sup> We excluded the state-level regressions for  $\phi^0$  because the Tobit model failed to converge to valid estimates for that specification.

**Table 17 – State-Level Tobit Regressions (RCA Cutoff = 1)**

Regressor	(1) $\phi^1$	(2) $\phi^1$	(3) $\phi^1$	(4) $\phi^1$	(5) $\phi^1$
$\theta^1$	266.180*** (10.3591)	267.571*** (10.3787)	272.940*** (7.4148)	274.292*** (7.4193)	266.261*** (10.3458)
inputs	92.877*** (9.5486)	94.201*** (9.5752)			93.229*** (9.5295)
laborint	13.587** (4.2798)	14.098** (4.2888)	19.832*** (3.1572)	19.710*** (3.1616)	13.781** (4.2723)
urban	148.944*** (8.5294)	148.595*** (8.5384)	166.156*** (5.9408)	166.306*** (5.9418)	122.797*** (10.4220)
numworkers	162.898*** (44.6597)		111.804*** (23.2372)		160.143*** (44.4848)
firmrev	-66.909 (66.9945)	4.214 (57.5821)	-105.663*** (20.0513)	-68.616*** (18.5084)	-65.790 (66.7217)
weightedurban					32.823*** (7.5253)
constant	-0.203** (0.0636)	-0.114* (0.0576)	-0.201*** (0.0237)	-0.129*** (0.0190)	-0.206** (0.0633)
N	2045199	2045199	11600937	11600937	2045199
Pseudo- $R^2$	0.2588	0.2540	0.0704	0.0693	0.2617

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.



**Table 18 – Firm-Level Tobit Regressions (RCA Cutoff = 0)**

Regressor	(1) $\phi^0$	(2) $\phi^0$	(3) $\phi^0$	(4) $\phi^0$	(5) $\phi^0$
$\theta^0$	966.331*** (74.9021)	966.786*** (74.9673)	1544.395*** (98.2073)	1548.667*** (98.4337)	966.377*** (74.9064)
inputs	112.833*** (6.1930)	113.216*** (6.1918)			112.837*** (6.1921)
laborint	17.874*** (1.5409)	18.025*** (1.5431)	37.020*** (2.0568)	37.015*** (2.0547)	17.877*** (1.5409)
urban	77.540*** (3.5129)	77.604*** (3.5155)	110.623*** (5.0838)	111.022*** (5.1024)	76.343*** (4.2063)
numworkers	50.804* (25.8557)		97.528*** (22.9448)		50.728* (25.8658)
firmrev	8.117 (13.3327)	33.020* (16.2007)	-19.534* (8.5558)	8.869 (8.2808)	8.149 (13.3470)
weightedurban					1.650 (2.9956)
constant	-0.252*** (0.0244)	-0.227*** (0.0176)	-0.408*** (0.0251)	-0.341*** (0.0166)	-0.253*** (0.0244)
N	2045199	2045199	11600937	11600937	2045199
Pseudo- $R^2$	0.286	0.285	0.133	0.131	0.286

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

**Table 19 – Firm-Level Tobit Regressions (RCA Cutoff = 1)**

Regressor	(1) $\phi^1$	(2) $\phi^1$	(3) $\phi^1$	(4) $\phi^1$	(5) $\phi^1$
$\theta^1$	962.833*** (75.0899)	963.366*** (75.1589)	1565.283*** (92.4176)	1569.778*** (92.6419)	962.875*** (75.0932)
inputs	114.108*** (6.2196)	114.518*** (6.2189)			114.111*** (6.2187)
laborint	17.961*** (1.5387)	18.124*** (1.5414)	37.549*** (2.0161)	37.543*** (2.0142)	17.965*** (1.5385)
urban	79.179*** (3.5108)	79.244*** (3.5137)	113.506*** (4.5380)	113.929*** (4.5548)	78.047*** (4.1701)
numworkers	55.261* (26.3226)		104.969*** (24.1138)		55.190* (26.3338)
firmrev	11.311 (14.1502)	38.927* (17.8076)	-17.590* (8.8343)	13.124 (8.7312)	11.338 (14.1655)
weightedurban					1.561 (2.9882)
constant	-0.263*** (0.0250)	-0.237*** (0.0192)	-0.425*** (0.0252)	-0.354*** (0.0158)	-0.263*** (0.0250)
N	2045199	2045199	11600937	11600937	2045199
Pseudo- $R^2$	0.284	0.283	0.132	0.130	0.284

Robust standard errors with multi-dimensional clustering in parentheses. All regressors are as defined in the paper and then divided by 1000. \* = significant at the 5% level. \*\* = significant at the 1% level. \*\*\* = significant at the 0.1% level.

Most of the variables are positive and significant across all specifications, as we might expect. The exceptions are firm revenue and the revenue-weighted urban similarity. The demand complementarities measure again dominates the model, with urban similarity, input similarity, and firm size similarity (as measured by number of workers) all roughly equal in terms of their effects. Firm revenue is inconclusive. It is negative in a few specifications, but is generally insignificant. The revenue-weighted urban share seems to matter only in the state-level specifications.

#### 4.1 Variance Partitioning

It is one matter to simply compare the magnitudes of coefficients in a table. However, we are also interested in knowing which measures have the largest contribution to the R-squared of the model. There are many different ways of measuring this, such as partial R-squared, or Shapley-Owen values. All of these measure fall into the broad category of techniques called Variance Partitioning.

Here we use a simple method of dividing up the R-squared across the regressors laid out in Peres-Neto et al. (2006)<sup>22</sup>, which looks at the marginal contribution that each variable makes to the R-squared when it is included in the model alongside all of the other variables. That is to say, it calculates,

$$r_X^2 = R_{Y|X,Z}^2 - R_{Y|Z}^2 \quad (15)$$

Where  $R_{Y|X,Z}^2$  is the R-squared of the full model, all variables included,  $R_{Y|Z}^2$  is the R-squared of the model with all variables *except*  $X$ , and  $r_X^2$  is the portion of the R-squared of the full model that is uniquely explained by  $X$ .<sup>23</sup> Another way of thinking of this is that  $r_X^2$  is the marginal contribution of  $X$

---

<sup>22</sup> Note that although we cite this paper because they have a clear explanation of the procedures and intuition for the method, the authors we cite did not invent variance partitioning. They merely extended a common method found in a textbook to a multi-variable setting. The earliest reference I can find to it in the literature is a textbook by Scherrer (1984), however because this textbook is not in English, and because the method is so common and so old, I was unable to track down the original inventor of the procedure.

<sup>23</sup> We acknowledge that this is not the optimal approach this this problem, and there are other alternatives in the literature. The preferred one is to compute the Shapley-Owen value for each variable. This is a measure based on a theorem from game theory that has been shown to have some nice properties, such as having all of the marginal contributions from each of the variables add to 1. Unfortunately, in our setting, calculating this value is not

after all other variables have been taken into account. Note that the marginal contributions will not necessarily add to the full R-squared, since part of the explanatory power of each variable might be shared with other variables in the model, or there may be explanatory power found in pairs of variables that is not there when the variables are included individually.

The results of this exercise are included in Table 20. What we find most interesting about the table is that the majority of explanatory power of the model seems to be coming not from the demand complementarities measure, but rather from the input similarity measure.<sup>24</sup>

**Table 20 – Contributions of Variables to R-squared**

Cutoff:	State	Firm	
	RCA=1	RCA=0	RCA=1
$\theta$	0.118	0.095	0.092
inputs	0.191	0.153	0.152
laborint	0.002	0.013	0.012
urban	0.033	0.03	0.031
numworkers	0.004	0.001	0.001
firmrev	0.001	0.00002	0.00003
weightedurban	0.002	0.00002	0.00002

Values based on the Tobit regressions found in the preceding tables.

computationally feasible. To implement it here, we would need to estimate our Tobit regression over  $2 * 8! = 80,640$  times, for *each* of our four settings (state and firm-level, and the two RCA cutoffs). Given how long it takes even one of these Tobit regressions to converge, we were forced to settle on a simpler measure, which while not as ideal as Shapley-Owen, still gives us a good picture of which variables are making the largest contribution to the R-squared.

<sup>24</sup> The most likely reason for this anomaly is the simple fact that because the input measure cannot be calculated for every product, its inclusion in the regression decreases the total number of observations and therefore makes it easier for the model to predict a larger portion of the sample.

What these results tell us is that although the demand complementarities seem to have the largest impact in terms of the magnitude of their effect (as measured by coefficient size), the variation in the proximity matrix seems to be most highly correlated with the variation in the input similarity measure. The next most highly correlated seems to be demand complementarities, which is a close second in each of the settings. The other variables, even though they often had positive and significant coefficients in the regression tables, seem to have negligible contributions to the R-squared for each of those models.

## **5 Conclusion**

This paper sought to investigate the foundations of Hidalgo et al.'s (2007) proximity measure, taking cues from the literature regarding what might cause certain products to be co-produced within the same firm or same geographical region. The strongest evidence seems to point towards demand complementarities, as measured by the probability that certain pairs of products are demanded together as intermediate inputs, as well as input similarities across products. The effect seems much larger at the firm level than at the state level, and it is robust to whether we consider product pairs for which a given producer has revealed comparative advantage, or whether we simply require the pairs to be produced (in any quantity) by the producer.

The other measures that one might expect to be meaningful based on the literature, are indeed so, but to a lesser degree. The typical number of workers in a firm that produces a product, the urban or rural environment, and the labor intensity of a product all have positive and significant effects on the probability that given product pairs will be co-produced. Surprisingly, the average revenue of a firm seemed to be a poor predictor. However, while these other variables are statistically significant in the regression models, they tend to have only a minor impact on the R-squared for those models. The

exceptions being the demand complementarity and input similarity measures, which dominated not only the regressions, but also the amount of explained variation.

## 6 References

- Balassa, Bela (1965). "Trade Liberalisation and 'Revealed' Comparative Advantage." *The Manchester School*. Volume 33, Issue 2. 99-123.
- Bernard, Andrew, Stephen Redding, and Peter Schott (2010). "Multiple-Product Firms and Product Switching." *American Economic Review*. 100:1, 70-97.
- Cameron, Colin, Jonah Gelbach, and Douglas Miller (2011). "Robust Inference with Multiway Clustering." *Journal of Business and Economic Statistics*. 29:2, 238-249.
- De Loecker, Jan (2011). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity." *Econometrica*. Vol 79, No. 5, 1407-1451.
- Duranton, Gilles and Diego Puga (2001). "Urban Diversity, Process Innovation, and the Life Cycle of Products." *The American Economic Review*. vol. 91, No. 5, 1454-1477.
- Ettlie, John and Albert Rubenstein (1987). "Firm Size and Product Innovation." *Journal of Product Innovation Management*. Vol. 4, Issue 2. 89-108.
- Feldman, Maryann and David Audretsch (1999). "Innovation in cities: Science-based diversity, specialization and localized competition." *European Economic Review*. 43. 409-429.
- Fieler, Ana (2011). "Nonhomotheticity and Bilateral Trade: Evidence and a Quantitative Explanation." *Econometrica*. Vol. 79, No. 4, 1069-1101.
- Flagge, Matthew and Ritam Chaurey (2015). "Firm-Product Linkages and the Evolution of Product Scope." Manuscript.
- Flagge, Matthew (2015). "Predictive Power of Proximity on Firm- and State-level production in India." Manuscript.
- Foster, Lucia, John Haltiwanger, and Chad Syverson (2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*. Vol. 98, No. 1, pp. 394-425.
- Gollop, Frank and James Monahan (1991). "A Generalized Index of Diversification: Trends in U.S. Manufacturing." *The Review of Economics and Statistics*. Vol. 73, No. 2, 318-330.
- Heckscher, Eli and Bertil Ohlin (1991). *Heckscher-Ohlin Trade Theory*. The MIT Press. Cambridge.
- Hidalgo, Klinger, Barabasi, and Hausmann (2007). "The Product Space Conditions the Development of Nations." *Science*. **317**, 482.
- Jacobs, Jane (1969). *The Economy of Cities*. New York: Random House.

Leamer, Edward (1984). *Sources of International Comparative Advantage: Theory and Evidence*. MIT Press. Cambridge, Massachusetts.

Penrose, Edith (1959). *The Theory of the Growth of the Firm*. Oxford: Basil Blackwell.

Peres-Neto, Pedro, Pierre Legendre, Stephane Dray, and Daniel Borcard (2006). "Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions." *Ecology*. Vol. 87, No. 10, pp. 2614-2625.

Papke, Leslie and Jeffrey Wooldridge (1996). "Econometric Methods for Fractional Response Variables with an Application to 401 (K) Plan Participation Rates." *Journal of Applied Econometrics*. Vol. 11, No. 6.

Poncet, Sandra and Felipe de Waldemar (Forthcoming). "Product Relatedness and Firm Exports in China." *World Bank Economic Review*.

Scherer, F. M. (1965). "Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions." *The American Economic Review*. Vol. 55, No. 5, 1097-1125.

Scherrer, B. (1984). *Biostatistique*. Gaëtan Morin, Chicoutimi.



# Chapter 3: Firm-Product Linkages and the Evolution of Product Scope\*

Matthew Flagge<sup>1</sup> and Ritam Chaurey<sup>2</sup>

<sup>1</sup>Columbia University

<sup>2</sup>SUNY Binghamton

February 2015

This paper provides a methodology for estimating the role of various measures of “distance” in firms’ product choice decisions. We model additions of new products by firms using a dynamic model in which firms must pay a one-time startup cost for adding new products to their production line. We allow this cost to be reduced if the firm already produces similar products, or shares characteristics with other firms already producing the product. We consider three characteristics along which firms may be considered “close” to a particular product: input similarity, physical distance to existing locations of production, and vertical connectedness. The set of potential product combinations is prohibitively large for standard estimation methods. Instead, we apply the method of moment inequalities developed by Pakes et al. (forthcoming) and Morales et al (2014). Results are heterogeneous across sectors, though physical distance seems to be of greatest importance. Counterfactuals in which we negate the benefits from certain proximity channels show that even in sectors where input similarity is important, physical proximity has a greater impact on the number of profitable products available to a firm.

---

\* We would like to thank Eric Verhoogen, Amit Khandelwal, Eduardo Morales, David Weinstein, Kate Ho, Jon Vogel, Don Davis, Peter Schott, Peter Neary, Jonathan Dingel, and Chris Conlon for helpful comments.

# 1 Introduction

How does a firm's product mix evolve? Consider the example of ITC Ltd., a large Indian conglomerate with over \$8 billion in revenue. This company started in 1910, producing tobacco, and entered the packing and printing business in 1925 as a form of backward integration. It began producing paperboard in 1979. In 1990 it began the exportation of agricultural commodities, which it describes as a leveraging of their agri-sourcing competency (ostensibly based on their existing ability to source wood and tobacco). They started producing notebooks in 2002, and later expanded to books, pens, pencils, and other stationary over the course of 2007-2009. They entered the food business with ready-to-eat meals in 2001, which their company website describes as "successfully blending multiple internal competencies."<sup>25</sup> They then progressed into confectionary and wheat flour (2002), biscuits (2003), and instant noodles (2010). Despite the incredible diversity of their current product basket, when viewed as a process of evolution, the expansion of their product scope did not happen in a random fashion. Rather, the firm appeared to follow certain natural paths, such as expanding into products that use their existing inputs or expertise, or pursuing vertical integration.

The nature of what a country's firms produce is not merely a subject of idle curiosity. There is theoretical literature that suggests that a country's products can matter for welfare. For instance, there can be learning, or spillovers across products (Matsuyama [1992], Harrison and Rodriguez-Clare [2010]). On the empirical side, Bernard, Jensen, and Schott (2006) find that the capital intensity of an industry's products can affect employment growth and the probability of plant death in the presence of international trade. Furthermore, Hidalgo et al. (2007) find the pairwise export correlations predict the development of future comparative advantage, which implies that countries whose exports are correlated with many products are more likely to develop comparative advantage in a broader range of

---

<sup>25</sup> <http://www.itcportal.com/about-itc/profile/history-and-evolution.aspx> (retrieved 9/16/2014)

products. These authors all suggest that both the type and diversity of the products produced by a country can have welfare effects for that country. Thus, a better understanding of the sequence in which products are added by firms can in turn give us a better understanding of the development path of a country, in terms of both product scope and welfare.

The question of what factors shape the evolution of a firm's product mix also relates to the active recent literature on multi-product firms in an international context. The existing literature offers two leading explanations for what might drive the sequence in which firms add products. Bernard, Redding, and Schott (2010) models the adding and dropping of products as the result of stochastic shocks to demand and firm-product productivity. Eckel and Neary (2010) employ a model in which firms have a core competency (lowest production cost) product, and firms add products in order of how similar they are to the core product. But the former model fails to account for the high frequency at which certain pairs of products are produced together, and the latter model is agnostic about what characteristics cause a product to be "near" or "far" from a firm's core competency.

Our previous chapters (Flagge 2015a, Flagge 2015b) examined a reduced-form approach to predicting which products firms will add, and what factors about the firms or products might be correlated with observed patterns of product co-production. However, these reduced-form approaches stop short of telling us *why* firms might be behaving in this way. What aspect of firm decision-making is affected by these factors and how big is their impact? To truly answer these questions, we need a structural model.

Our paper develops a methodology that allows us to estimate the costs that firms face in transitioning to new products, and calculate how those costs vary based on certain measures of "distance" between firms and products. We consider three such measures within this paper: 1)

Overlapping inputs, 2) Physical proximity of the factory to other locations where the product is produced, 3) upstream/downstream connectedness via input-output linkages.

Determining the topology of the product landscape is a non-trivial undertaking. Modelling a decision as complex as product choice would be difficult in a discrete-choice setting. The size of the choice set is very large, and the problem would be computationally infeasible even if firms' information sets were known. We circumvent these difficulties by using a novel econometric technique called moment inequalities, developed by Pakes, Porter, Ho, and Ishii (forthcoming) [henceforth, PPHI]. The method relies on a "revealed preferences" assumption. Rather than trying to explicitly model firms' choices, we observe their actions and assume they are at least weakly more profitable (on average) than their other possible choices.<sup>26</sup> This allows us to derive an inequality condition where on one side are the expected profits for engaging in the chosen action, and on the other are profits from a potential counterfactual choice. Each of these profit terms is a function of parameters defined in a theoretical model, and these inequalities allow us to find upper and lower bounds on the parameters (i.e. the highest and lowest values of the parameters that are consistent with the inequalities derived from the firm choices).

The theoretical and empirical framework for our analysis closely follows Morales, Sheu, and Zahler (2014) [henceforth, MSZ], a structural gravity model with a dynamic component to capture how firms' costs of entry into a new market might depend on their prior entry choices. MSZ studies firm entry into country markets, which are distanced from the firm in physical space. We adapt their model to study firm entry into product markets, where each new product has a distance from the firm within a "characteristic space." This model is able to capture the dynamic component of firm choice,

---

<sup>26</sup> The full assumptions we make on firm behavior are made explicit in Section 4.3 of the paper. For the time being, it's worth noting that the assumptions we need are consistent with, but substantially weaker than, perfect rationality.

incorporating the connections that potential new markets have to firms' existing abilities. In the model, firms choose whether to add new products, and which products to add, out of a universe of possible products. Each firm-product pair has a stream of projected revenue that it can offer the firm, but entry is deterred by startup costs the firm must incur to begin production of a particular product. These startup costs depend on whether the firm is "close" to the new product, along the three dimensions enumerated earlier.

The data we use come from India's Annual Survey of Industries, a factory-level dataset that includes inputs, outputs, and physical location, among many other characteristics. The data are an unbalanced panel with yearly observations, chosen because it allows us to observe adding of products by firms in an emerging markets setting.

Our results are bounds on the costs of transitioning into new products. We estimate these costs separately by sector, and results are heterogeneous across sectors. In general, the physical proximity measure seemed to perform the best out of the three, across all sectors. Counterfactual exercises in which we calculate the number of profitable products that would be available to firms if we nullified the effects from one of the distance measures support this. Removing the cost benefits received from physical proximity has the greatest impact on the number of potentially profitable products firms' have available.

The paper will proceed as follows. Section 2 discusses the dataset. Section 3 offers some preliminary evidence from our data. Section 4 describes the model. Section 5 outlines the procedure by which the model is estimated. Section 6 provides the results. Section 7 performs some supplementary analyses, such as simulation of product entry by firms and counterfactuals. Section 8 concludes.

## 2 Data

The primary dataset we use is the panel portion of the *Annual Survey of Industries (ASI)* from India. This is an unbalanced panel spanning the years 1999-2008. The data are a representative sample of all factories with 20 or more employees without power, and 10 or more employees if the factories have power.

The standard panel dataset for the ASI includes (among other items), land, buildings, physical plant, workers (male, female, child, managerial, and contractors), wages, material inputs and their costs, fuel and electricity usage, and outputs and their associated revenues.

The data also have an associated cross-sectional version, which lacks unique identifiers for factories. We merged the cross-section with the panel in order to observe plant location at the district level, as well as the number of plants per firm.

In selecting firms for inclusion in our study, we dropped all factories that<sup>27</sup>:

1. Do not appear in at least two consecutive years, or
2. Did not fill out one of the blocks of the survey required for our analysis (inputs, outputs, employment, expenses), or
3. Provided only aggregate output data, or
4. Classified all outputs as “miscellaneous.”

Table 21 presents some summary statistics for the data. As we can see, almost all factories in the data belong to single-factory firms. Thus, in this paper, we will use the terms *factory* and *firm* synonymously. The large proportion of single-factory firms is a useful feature of our data, because it implies our estimates will be informative for understanding firm strategy, as opposed to being based on

---

<sup>27</sup> We also performed a robustness check in which we excluded all factories that were part of a collection of factories belonging to the same owner. This did not have any qualitative impact on our results.

incomplete information about products being transferred from one factory to another within the same firm. As a note, single-factory firms tend to be smaller than multi-factory firms, and within our data they represent a less than proportional share of output, but they nevertheless represent a non-trivial portion of the economic output counted by our dataset (84% of all revenues).

We can also see that products were added in 37% of the firm-years in the data. Having such a large number of observations in which products are added will be helpful for our estimation procedure, which relies on analyzing firm behavior, such as adding products.

Other observations from the table are that the firms in the dataset use a rich set of inputs, which will be helpful in analyzing how their input mix affects product choice. The average revenue per product line is included in the table to give readers a perspective on the magnitude of our coefficients when we provide our estimates later in the paper. Table 22 provides the mean and median revenues separated out by industry, given in thousands of 1982 dollars.

**Table 21 - Summary Statistics**

	Mean (Std. Dev)	Observations (firm-years)
Number of products	2.16 (1.85)	192345
% that added products*	0.37	179972
Number of products added**	1.54 (1.00)	66464
Revenue per product line***	443378.1 (3605142)	192345
% Single-factory firms	0.94	209857
% of revenue from single-factory firms	0.84	192586
Number of inputs (indigenous)	4.81 (3.15)	191085
Number of inputs (imported)	10.75 (3.28)	197166

\* Among single-factory firms it is 36%

\*\* Conditional on adding a product

\*\*\* Expressed in 1982 rupees

**Table 22 – Revenue per product line by industry (1000s of 1982 dollars)**

	Animal, Vegetable, Forestry	Ores, minerals, gas, electricity	Chemicals	Rubber, plastic, leather
mean	28	189	53	36
median	2	1	4	3
obs.	87,402	20,578	50,941	22,897
	Wood, cork, paper	Textiles	Metals, Machinery	Railways, ships, transport
mean	18	37	49	176
median	1	6	2	6
obs.	21,474	44,799	108,765	9,825



### 3 Preliminary Evidence

Here we will present some reduced form evidence to show that the cluster correlations we are looking for exist within our dataset, and will try to convince the readers that the explanations offered by the standard models do not adequately explain these clusters.

Table 23 displays the conditional probabilities that a firm whose primary product (defined as the product generating the most revenue for that firm) is in the row sector in period  $t$  will start producing a product in the column sector in period  $t+1$ .<sup>28</sup> The colors in the table merely highlight the relative magnitude of the matrix elements and are not meant to convey any additional information beyond what is already contained within the elements of the table.

As can be seen from the table, firms have a tendency to add products to their basket from within their own sector. However, there are also a sizeable number of firms that add products from other sectors. It is worth noting that the zeros in the table are “rounded zeros.” That is, those elements in the table are very small, but not identically zero. We can deduce from this that path of a firm through the product space is potentially very complicated, and it would be difficult to feasibly model this decision and the choice set in a discrete-choice framework, thus necessitating the use of moment inequalities.

The pattern observed in Table 23 persists even if we move to a greater level of disaggregation and observe a single sector. Firms continue to add products predominantly along the diagonal, indicating a tendency towards new products that are similar to ones they already produce.

---

<sup>28</sup> Rows in the table do not add to 1 due to the presence of some firms adding multiple products in the same period.

**Table 23**

Main sector in previous year		Conditional probability of adding product in a sector								
		1	2	3	4	5	6	7	8	9
<b>1</b>	<b>Animal, vegetable, forestry</b>	0.9	0.02	0.06	0	0	0.01	0.01	0	0
<b>2</b>	<b>Ores, minerals, gas electricity</b>	0.01	0.81	0.06	0.01	0	0	0.06	0	0.05
<b>3</b>	<b>Chemicals</b>	0.06	0.05	0.8	0.03	0.01	0.01	0.03	0	0.02
<b>4</b>	<b>Rubber, plastic, leather</b>	0.01	0	0.04	0.69	0.02	0.08	0.1	0.03	0.02
<b>5</b>	<b>Wood, cork, paper</b>	0.01	0	0.02	0.03	0.84	0.01	0.05	0	0.03
<b>6</b>	<b>Textiles</b>	0.02	0	0.01	0.04	0.01	0.92	0.01	0	0
<b>7</b>	<b>Metals and machinery</b>	0	0.02	0.02	0.04	0.01	0.01	0.83	0.05	0.03
<b>8</b>	<b>Railways, ships, other transport</b>	0	0	0	0.07	0.01	0	0.48	0.42	0.02
<b>9</b>	<b>Other manuf. articles and services</b>	0	0.07	0.02	0.04	0.03	0.02	0.19	0.01	0.62

Table 24 shows a similar conditional probability matrix for three-digit product categories within sector 77 (electrical machinery). As we indicated, firms tend to add new products along the diagonal. However, there are also substantial product additions in “close” categories. For instance, those firms manufacturing domestic and office equipment (777) are likely to add electrical machinery (771). Those firms making switchgear and control panels (773) add measuring and controlling instruments (775).

**Table 24 – Electrical and Electronic Machinery or Equipment**

Main sector in previous year		Conditional probability of adding product in a sector								
		771	772	773	774	775	776	777	778	779
<b>771</b>	<b>Electrical Machinery</b>	0.20	0.03	0.01	0.02	0.03	0.01	0.04	0.03	0.03
<b>772</b>	<b>Motors, generators, transformers</b>	0.04	0.22	0.05	0.02	0.05	0.01	0.00	0.03	0.05
<b>773</b>	<b>Switchgear, control panels</b>	0.02	0.06	0.27	0.05	0.09	0.00	0.00	0.02	0.07
<b>774</b>	<b>Lamps, filaments, electrodes</b>	0.01	0.01	0.03	0.37	0.02	0.00	0.01	0.01	0.05
<b>775</b>	<b>Measuring/controlling instruments</b>	0.02	0.06	0.08	0.02	0.24	0.01	0.00	0.01	0.07
<b>776</b>	<b>Batteries and cells</b>	0.03	0.04	0.00	0.03	0.03	0.51	0.00	0.00	0.02
<b>777</b>	<b>Domestic and office equipment</b>	0.10	0.02	0.02	0.04	0.00	0.00	0.17	0.04	0.04
<b>778</b>	<b>Electromagnetic equipment</b>	0.03	0.03	0.04	0.04	0.02	0.00	0.00	0.25	0.06
<b>779</b>	<b>Electrical equipment, n.e.c.</b>	0.02	0.05	0.07	0.08	0.07	0.00	0.02	0.02	0.12

## 4 Theoretical Framework

This section outlines the theoretical framework we use for our estimation. In a study of the connections between products, one might imagine that product linkages can exist on both the supply and demand sides of the market. For this exercise, we exclude the possibility of demand-side linkages, and focus only on supply-side features of products.<sup>29</sup>

The model we use is a modification of the model found in MSZ, but adapted to model the entry of firms into product markets rather than into locational markets. While the use of this type of model to study this type of problem may be unprecedented, the basic intuition underlying it applies to our situation as well as it applies to the problem of international trade. In their model, exporters select destination markets, favoring larger markets, and disfavoring markets that are further away. In our adaptation, the process is the same, except the destination markets are product lines rather than physical locations, and the “distance” between the firm and the destination is a startup cost for that product line, rather than the trade costs associated with physical distance.

### 4.1 Demand

Demand is modeled in the style of Dixit and Stiglitz (1977). There is a representative consumer with CES utility over varieties  $i$  in a given product category  $j$ . The consumer has separable utilities over product categories, with the utility in any period  $t$  from category  $j$  given by:

$$Q_{jt} = \left[ \int_{i \in A_{jt}} q_{ijt}^{\frac{\eta_j - 1}{\eta_j}} di \right]^{\frac{\eta_j}{\eta_j - 1}} \quad \eta_j > 1 \quad (16)$$

---

<sup>29</sup> We admit this is a strong assumption. However, it is made primarily due to data constraints, as opposed to prior beliefs by the authors regarding the drivers of firm product choice. We are not currently aware of a dataset that allows us to observe demand side linkages and connect them to our current list of firms and products. Existing data that we are aware of uses different product classifications than those found in the ASI, and we have not found a concordance to match the two. It may be possible to relax this assumption in future versions of the paper.

Where  $A_{jt}$  is the set of available varieties,  $\eta_j$  is the elasticity of substitution for products of type  $j$ , and  $q_{ijt}$  is the consumption of variety  $i$  in time  $t$ .

The demand for varieties that emerge out of this utility function is:

$$q_{ijt} = \frac{p_{ijt}^{-\eta_j}}{P_{jt}^{1-\eta_j}} C_{jt} \quad (17)$$

Where  $P_{jt}$  is a price index given by:

$$P_{jt} = \left[ \int_{i \in A_{jt}} p_{ijt}^{1-\eta_j} di \right]^{\frac{1}{1-\eta_j}} \quad (18)$$

In the above index,  $p_{ijt}$  is the price of a given variety and  $C_{jt}$  is the total consumption of all products of type  $j$ .

## 4.2 Supply

Firms in the model must choose whether they will produce a variety in a given product category

$j$ . Firms that choose to produce will face three types of costs:

1. Marginal costs:  $mc_{fjt}$
2. Fixed costs:  $fc_j$
3. Product startup costs:  $sc_{fjt}(b_{t-1})$

We will explain each of these elements in turn.

### 4.2.1 Marginal Costs

Similar to Goldberg, Khandelwal, Pavcnik, and Topalova (2010), we give firms a Cobb-Douglas production function:

$$q_{ijt} = (\beta_{ft}^{mc})^{-1} L_{fjt}^{\beta_L^{mc}} IC_{fjt}^{\beta_{IC}^{mc}} \quad (19)$$

Where  $L_{fjt}$  is the labor assigned by firm  $f$  to product  $j$  in period  $t$ , and  $IC_{fjt}$  is the basket of intermediate inputs used in product  $j$ , and  $\beta_L^{mc} + \beta_{IC}^{mc} = 1$ .

This yields a log-linear form for marginal costs, as follows:

$$\ln(mc_{fjt}) = \beta_{ft}^{mc} + \beta_L^{mc} \ln(PL_j) + \beta_{IC}^{mc} \ln(PIC_{fjt}) + \epsilon_{fjt}^{mc} \quad (20)$$

Where  $PL_j$  and  $PIC_{fjt}$  are the price of labor and the price of the intermediate input basket respectively, and  $\epsilon_{fjt}^{mc}$  is an error term. Please see the appendix, section 1, for details on the calculation of each of these terms.

#### 4.2.2 Fixed Costs

Fixed costs reflect costs the firm incurs every year it produces product  $j$ , regardless of the quantity produced. We set fixed costs to be static for every product, but allow them to vary across industries.<sup>30</sup> We denote the industry for product  $j$  as  $J$ , where by industry we mean the 1-digit product classification associated with product  $j$ .

$$fC_{fjt} = \mu_J^{fc} + \epsilon_{fjt}^{fc} \quad (21)$$

---

<sup>30</sup> Previous versions of our estimation included more parameters, including labor, capital, or labor intensity. However, these were found not to have a significant effect. In MSZ, they include many of the terms from the startup costs in the fixed cost equation as well. However, they are able to do this because there exists static versions of the startup costs in their framework. Specifically, they can look at the “distance” between Chile and another country (which is static), as opposed to the distance between a firm and another country (which is dynamic). However, in our framework, all of the distance measures are inherently dynamic. There are no static country-level versions to incorporate. Thus, in order to stay true to the nature of their model, in which the dynamics only appear in the startup costs, we avoid including the distance terms in our fixed cost.

### 4.2.3 Product Startup Costs

These are analogous to the sunk costs in MSZ, and are paid by firms that are producing  $j$  in a given period, but did not produce it in the previous period. They reflect the initial costs of setting up a new production line, and can be diminished if a product is “closer” to a firm along a certain distance measure. For instance, if a new product shares inputs with one or more of the firm’s existing products, this diminishes or eliminates the search cost for the firm to find a supplier of these inputs, and potentially eliminates a learning cost associated with discerning how to use those inputs effectively.

The startup costs in period  $t$  are defined to be a function of the firm’s “basket” in the previous period, which we denote as  $b_{t-1}$ . The basket is the collection of characteristics of the firm in any given period. It is, most notably, the whole range of products produced by the firm in that period, but can also include less tangible characteristics (such as proximity of the firm to production locations of other products). By defining the startup costs as being a function of  $b_{t-1}$  (as opposed to  $b_t$ ), we are restricting the costs the firm has to pay to begin production of a new product to be determined by characteristics of the firm *prior* to making the decision to produce.

The startup costs are modeled as follows:

$$SC_{fjb_{t-1}t} = \mu_j^{SC} - e_j^{SC}(b_{t-1}) + \epsilon_{fjt}^{SC} \quad (22)$$

$$e_j^{SC}(b_{t-1}) = \zeta_1^{SC} \phi_j^1(b_{t-1}) + \zeta_2^{SC} \phi_j^2(b_{t-1}) + \zeta_3^{SC} \phi_j^3(b_{t-1})$$

In the above equations, the  $\phi_j$  are proximity measures, ranging from 0 to 1, where 1 indicates a destination product  $j$  is considered “close” to a firm along a certain measure of distance. We have three such distance measures we are considering in this paper, which we will explain in turn.

#### 4.2.3.1 Distance Measure 1: Similarity of Input Cost Shares

This distance measure corresponds to the variable  $\phi_j^1(b_{t-1})$  in the equation for product startup costs. We use Kugler and Verhoogen's (2012) modified Gollop and Monahan (1991) measure of horizontal differentiation. We use it to capture whether a firm  $f$ , seeking to produce product  $j$  uses similar inputs to other firms already producing  $j$ . The index ranges from 0 to 1, where 0 represents completely identical inputs (measured in terms of cost share), and 1 represents completely dissimilar inputs. The index is calculated as follows, for any two firms  $f$  and  $f'$ :

$$\sigma_{ff'} = \left( \sum_m \frac{|w_{fm} - w_{f'm}|}{2} \right)^{\frac{1}{2}} \quad (23)$$

Where  $w_{hm}$  is the cost share of input  $m$  into firm  $h$ .

Having calculated  $\sigma_{ff'}$  for every pair of firms, we define the distance from a firm to a product to be the minimum of the distances to the firms already producing the desired product. After computing this distance index, we convert this distance to a proximity,  $\phi^1$ , which in this case merely requires reversing the distance. More precisely:

$$\phi_{fj}^1(b_{t-1}) = \left| \left( \min_{f' \in \mathcal{F}_{j,t-1}} \sigma_{ff'} \right) - 1 \right| \quad (24)$$

Where  $\mathcal{F}_{j,t-1}$  is the set of all firms already producing  $j$  in  $t-1$ .<sup>31</sup> If  $\mathcal{F}_{j,t-1}$  is the empty set, then we say  $\phi^1$  is undefined. The  $|\cdot|$  is the absolute value operator.

---

<sup>31</sup> It is worth noting that although we only use 44,022 firms to find observations for the moments (see the Data section of the paper for a discussion of this), we use all available firms in the dataset (over 100,000) to compute the modified Gollop and Monahan distance measure. This was to avoid the possibility that a firm producing  $j$  and having very similar inputs to a firm  $f$  would be excluded from the calculation because it did not satisfy the criteria needed in order to be used for the moment inequality estimation.

By including this measure in our estimation, we hope to capture some of the costs that firms must incur in order to add new inputs to their production lines. These could include costs such as finding suppliers, learning about new inputs, purchasing machines to process these inputs, training employees to use the new inputs, etc.

#### 4.2.3.2 Distance Measure 2: Physical Distance

Our second distance measure gives the physical distance between a selected firm  $f$  and the nearest firm already producing its destination product  $j$ . We do not have the exact location of firms in the data, but we do know a firm's district, out of 619 districts in India that were indexed by the Ministry of Statistics and Programme Implementation (MOSPI). See Appendix section 2 for a discussion of how districts were mapped to firms, as well as further details on the distance calculation.

#### 4.2.3.3 Distance Measure 3: Upstream/Downstream Connectedness

Our third type of distance measures how connected products are via upstream or downstream linkages, as determined by our input-output table. This is distinct from Measure 1 (input similarity). For two products,  $i$  and  $j$ , Measure 1 tells us whether  $i$  and  $j$  share similar inputs, whereas Measure 3 tells us whether  $i$  is used as an input in  $j$  (or vice versa). The formula we use to represent this is as follows:

$$\phi_{fj}^3(b_{t-1}) = \max_{i \in b_{t-1}} (\max\{w_{ij}, w_{ji}\}) \quad (25)$$

where  $w_{ij}$  is the cost share of input  $i$  into product  $j$ .

Because this is a measure of distance, we want it to be symmetric. Thus, we view the use of  $i$  in  $j$  and the use of  $j$  in  $i$  equivalently.  $\max\{w_{ij}, w_{ji}\}$  gives us the defined proximity between two products,



and after computing this for every product pair, the proximity of the firm to the given product  $j$  is simply the distance of the closest product to  $j$  found within the firm's basket in the previous period,  $b_{t-1}$ .

This measure of proximity varies between 0 and 1, with  $\phi^3 = 0$  if none of the firm's products use product  $j$  as an input, nor are used in the production of  $j$ . On the other hand,  $\phi^3 = 1$  if the firm possesses at least one product whose *only* input is product  $j$  (or alternatively, if any of the firm's products are the only input *in* product  $j$ ).

### 4.3 Firms' Optimal Behavior

The above theoretical framework yields the following profit function for firms:

$$\pi_{ft}(b_t|b_{t-1}) = \sum_{j \in b_t} \pi_{fjt}(b_{t-1})$$

$$\pi_{fjt}(b_{t-1}) = v_{fjt} - f c_{fjt} - \mathbb{I}\{j \notin b_{t-1}\} s c_{fjt}(b_{t-1})$$
(26)

Intuitively, a firm's total profit is equal to the sum of the profits from its individual product lines.  $\mathbb{I}\{\cdot\}$  is an indicator function, and  $v_{fjt}$  is the gross value of producing  $j$  to firm  $f$  in period  $t$ , as calculated from the demand function. The marginal costs are incorporated into the calculation of  $v_{fjt}$ , thus they do not appear separately in the profit function. We will explain the estimation of  $v_{fjt}$  in the section on the first stage estimation, to follow shortly.

As in MSZ, firms in this model solve a two-stage problem to determine which product lines to enter. The first stage is static, in which the firm looks at the universe of all products, and calculates the

expected gross profits from entering into each of those products<sup>32</sup>. The second stage is dynamic, in which the firm chooses which products to produce, factoring in the fixed costs and startup costs.

There are a number of assumptions that need to be made about firm behavior in order to estimate this model. We borrow these assumptions from MSZ, and modify them only to fit the notation found in this paper.

**Assumption 1:** *Let us denote by  $b_1^T = \{b_1, b_2, \dots, b_T\}$  the observed sequence of baskets chosen by any given firm  $f$  between periods 1 and  $T$ . Given a sequence of information sets for firm  $f$  at different time periods,  $\{J_{f_t}, J_{f_{t+1}}, \dots\}$ , a sequence of choice sets from which firm  $f$  picks its preferred basket,  $\{B_{f_t}, B_{f_{t+1}}, \dots\}$ , and a particular conditional expectation function  $\mathbb{E}[\cdot]$  capturing its subjective expectations, we assume:*

$$b_t = \operatorname{argmax}_{o_t \in B_{f_t}} \mathbb{E}[\Pi_{f_t}(o_t | b_{t-1}) | J_{f_t}] \quad \forall t = 1, 2, \dots, T$$

Where

$$\Pi_{f_t}(o_t | b_{t-1}) = \pi_{f_t}(o_t | b_{t-1}) + \delta \pi_{f_{t+1}}(o_{t+1} | o_t) + \omega_{f_{o_{t+1}t+2}} \quad (27)$$

The term  $\omega_{f_{o_{t+1}t+2}}$  is any arbitrary function that satisfies:

$$(\omega_{f_{o_{t+1}t+2}} \perp o_t) | o_{t+1} \quad (28)$$

And the basket  $o_{t+1}$  is defined as the optimal basket that would be chosen at period  $t + 1$  if the basket  $o_t$  was chosen at period  $t$ :

---

<sup>32</sup> We define “gross” here to mean profits before subtracting fixed costs and startup costs. Gross profits *do* take into account marginal costs.

$$o_{t+1} = \operatorname{argmax}_{\sigma_{t+1} \in \mathcal{B}_{f,t+1}} \mathbb{E}[\Pi_{f,t+1}(\sigma_{t+1}|o_t)|\mathcal{J}_{f,t+1}] \quad (29)$$

Assumption 1 imposes that the basket actually chosen by the firm must be the one that maximizes its value function ( $\Pi_{f,t}$ ) in expectation, where the expectations of the firm are based on  $\mathcal{J}_{f,t}$ , the information set of the firm in the period in which it is making the decision. It also imposes that the firm takes into account the effect of its decisions on future profits at least one period ahead. Note, this is still consistent with firms that are perfectly forward looking (for instance, if  $\omega_{fo_{t+1}t+2}$  is the discounted stream of all future profits).

Equation (28) imposes that the basket choice in period  $t$  does not affect firm profits beyond period  $t + 1$ , *except* through its effect on the basket choice the firm makes at  $t + 1$ . This is because the startup costs the firm must pay in period  $t$  only depend on the basket in period  $t - 1$ , and not in any prior periods. Furthermore, the firm internalizes that its choice in period  $t + 1$  is going to be the result of an analogous optimization problem to the one it solved in period  $t$  (see equation (29)).

Assumption 1 does not impose any constraints on the expectation functions of the firms, the firms' information sets, nor on the choice sets<sup>33</sup>, all of which may differ by firm, and the latter two of which may differ by period.

Assumption 1 implies the following:

**Corollary 1:**<sup>34</sup> *If Assumption 1 holds, and  $b'_t \in \mathcal{B}_{f,t}$ , then:*

---

<sup>33</sup> In finding observations for the estimation of the moment inequalities, we do assume a certain minimum size for the choice sets in order to generate our perturbations. The types of one-period deviations we consider are: 1) Beginning production of a product one period earlier than was actually chosen; 2) Delaying production of a product for one period; 3) Choosing production of some alternate product in lieu of a product the firm actually chose; 4) Choosing production of a product in lieu of non-production; and 5) Choosing non-production of a product in lieu of production. Thus, we require the choice set to include the firms' actual choices, as well as a small space of perturbations around those choices. This is nowhere near the size of the space of all possible firm choices, although our framework does not exclude the possibility that firms are using that space.

$$\mathbb{E}[\pi_{f_t}(b_t|b_{t-1}) + \delta\pi_{f_{t+1}}(o_{t+1}|b_t)|\mathcal{J}_{f_t}] \geq \mathbb{E}[\pi_{f_t}(b'_t|b_{t-1}) + \delta\pi_{f_{t+1}}(o_{t+1}|b'_t)|\mathcal{J}_{f_t}] \quad (30)$$

Where

$$o_{t+1} = \operatorname{argmax}_{\sigma_{t+1} \in \mathcal{B}_{f_{t+1}}} \mathbb{E}[\Pi_{f_{t+1}}(\sigma_{t+1}|o_t)|\mathcal{J}_{f_{t+1}}]$$

This corollary is used to derive observations for the moment inequalities, based on Assumption

1. It states that the observed basket choice by the firm must be at least weakly more profitable (in expectation) than any other basket that was in the firm's choice set.

Assumption 1 and its associated corollary allow us to apply an analogue of Euler's perturbation method with one-period deviations to the analysis of single-agent dynamic discrete choice problems, like the one we are analyzing.<sup>35</sup> This lets us obtain our estimates without the need to compute the fixed point for the value function, which would be infeasible in a problem of this size.

Each of the  $\pi$  functions expressed in equation (30) is a function of the parameters we are seeking to estimate. The estimation method then consists of solving a linear programming problem to find the values of those parameters that are consistent with a set of inequalities of a form analogous to equation (30). As one might surmise, inequalities with fewer terms lead to less ambiguity about the acceptable values of the parameters.<sup>36</sup> It is thus desirable to generate simpler inequalities when

---

<sup>34</sup> This corollary to Assumption 1 is equivalent to "Proposition 1" in MSZ, and is proved in the appendix of their paper.

<sup>35</sup> See Pakes, Porter, Ho, and Ishii (2011) for further details.

<sup>36</sup> As an example of this, consider the following two sets of inequalities:

$$\left. \begin{array}{l} \{2 \leq x \leq 4\} \\ \{1 \leq y \leq 2\} \end{array} \right\} \quad \left. \begin{array}{l} \{3 \leq x + y \leq 6\} \\ \{1 \leq y \leq 2\} \end{array} \right\}$$

The first set generates a smaller range of acceptable values for  $x$ : [2,4] vs [1,5]. Because  $x$  appears with  $y$  in the second set's inequality, any ambiguity in the true value of  $y$  propagates into  $x$ .

possible. This end is aided by the use of one-period deviations. Equation (28) allows us to ignore the terms of the profit function beyond period  $t + 1$  whenever we use a one-period deviation in period  $t$  to generate an inequality. Since (28) guarantees the profit beyond  $t + 1$  is the same in both the actual and counterfactual scenarios, the profit terms past  $t + 1$  simply cancel out, leading to inequalities of the sort found in equation (30).

Our procedure also requires some assumptions about the firms' choice sets and information sets. The constraints that we impose on the choice sets are laid out in Assumption 2:

**Assumption 2:** *Let us denote by  $\mathcal{B}_{ft}$  the choice set of  $f$  at  $t$ , and by  $b_t$  its optimal basket. Then:*

$$(b_t, \{\bar{b}_{jt}; \forall j\}, \{\bar{b}_{jj't}; \forall j, j'\}) \in \mathcal{B}_{ft}$$

where  $\bar{b}_{jt}$  is the basket that results from modifying the value corresponding to  $j$  in  $b_t$ , and  $\bar{b}_{jj't}$  is the basket that results from exchanging elements  $j$  and  $j'$  in  $b_t$

This assumption requires the choice set of any given firm to include, at the very least, the actual observed choice of the firm ( $b_t$ ), and a small number of perturbations around it. Requiring  $\bar{b}_{jt}$  to be in the choice set means that a firm could have chosen to produce either one more, or one less product than it actually chose to produce. Requiring  $\bar{b}_{jj't}$  to be in the choice set means the firm could have produced some other product, instead of one of the products it actually chose to produce.

Note that Assumption 2 is consistent with a firm's choice set including the whole universe of possible product combinations, but it does not require the choice set to be so large. Rather, it only imposes certain minimum requirements on the choice set.

We further have Assumption 3, imposing the minimum necessary contents of the firms' information sets:

**Assumption 3:** *Let us denote by  $\mathcal{I}_{ft}$  the information set of  $f$  at  $t$ . Then,*

$$Z_{ft} \in \mathcal{I}_{ft}$$

where  $Z_{ft} = \{Z_{fjt}; \forall j \in \mathcal{B}_{ft}\}$ , and  $Z_{fjt}$  includes  $b_{t-1}$ ,  $\mu_j^{fc}$ ,  $\mu_j^{sc}$ , and all of the covariates determining  $r_{fjt}$  and  $e_j^{sc}$ .

So at the time in which the firm must choose its basket for the current period, Assumption 3 requires the firm to know its basket in the previous period ( $b_{t-1}$ ), the determinants of the expected gross revenue it would receive ( $r_{fjt}$ ),<sup>37</sup> and the determinants of the fixed and startup costs ( $\mu_j^{fc}$ ,  $\mu_j^{sc}$ ,  $e_j^{sc}$ ) that it would face if it were to produce any given product under consideration (less any  $\epsilon$  error terms included in the equations for those costs).

## 5 Estimation

Estimation proceeds in two stages, mirroring the two-stage optimization problem of the firm. In the first stage, we compute the expected gross profits for each firm of entering each product market. In the second stage, we employ moment inequalities using the firms' observed choices to estimate the parameters of interest ( $\mu$  and  $\zeta$ ). This two-stage estimation allows us to generate moment inequalities that are linear in the parameters of interest<sup>38</sup>, thus avoiding the added computational difficulty of estimating with non-linear moments.

### 5.1 First Stage

We use the first stage to find point estimates for the parameter vector  $\beta$  found in equation (20). The subsequent estimates of the  $\mu$  and  $\zeta$  parameters in the model<sup>39</sup> will depend on this  $\beta$ . A difficulty

---

<sup>37</sup> We have not introduced this term yet, but we will be discussing it shortly, at the beginning of section 5.

<sup>38</sup> As will be shown, the moments are linear in all parameters except  $\beta$ , in which they are log-linear.

<sup>39</sup> See equations (21) and (22) for  $\mu$  and  $\zeta$ .

arises because (20) is an equation for marginal costs, which are typically unobserved. However, from the Dixit-Stiglitz demand system in our model, we can calculate the gross revenue a firm could expect from producing  $j$  in period  $t$ :

$$r_{fjt} = \left( \frac{\eta_j}{\eta_j - 1} \frac{mc_{fjt}(\beta)}{P_{jt}} \right)^{1-\eta_j} C_{jt} \quad (31)$$

This equation is log-linear, so we can take the log of (31), collect all the observable variables into a vector that we shall call  $z_{fjt}$ , and estimate the  $\beta$ 's with the following regression:

$$\ln(r_{fjt}) = \beta z_{fjt} + (1 - \eta_j) \epsilon_{fjt}^{mc} \quad (32)$$

Where  $z_{fjt}$  includes all observable variables in equation (20),  $\eta_j$  is taken as given, and  $\epsilon_{fjt}^{mc}$  is assumed to be independent of all variables included in  $z_{fjt}$ . We use a power function of the market size (total sales of product  $j$ ) to proxy for the  $P_{jt}^{\eta_j-1} C_{jt}$  in equation (31), and include firm-year fixed effects.

We then take the predicted values from this regression and convert them to levels— $\exp(\hat{\beta} z_{fjt})$ —to get preliminary predictions for the revenue. However, as pointed out by Santos Silva and Teneyro (2006), estimating log-linear models with OLS can be biased due to Jensen's Inequality. As an ad hoc way of addressing this potential bias, we take the observed revenues and regress them on the predictions, with no constant:

$$r_{fjt} = \alpha \exp(\hat{\beta} z_{fjt}) + \epsilon_{fjt}^r \quad (33)$$

The predicted  $\hat{\alpha}$  from this regression is then used to generate our final predictions for the revenue, as follows:

$$\hat{v}_{fjt} = v_{fjt}(\hat{\beta}) = \frac{1}{\eta_j} \hat{r}_{fjt} = \frac{1}{\eta_j} \hat{\alpha} \exp(\hat{\beta} z_{fjt}) \quad (34)$$

Because the elasticities of substitution  $\eta_j$  are not identified in this framework, we use the values calculated by Broda, Greenfield, and Weinstein (2006)<sup>40</sup>. Denote the error in our estimate of  $\hat{v}_{fjt}$  as  $\epsilon_{fjt}^v$ .

As a robustness check for our predictions, we also performed the first stage regression in levels (as opposed to performing it in logs, and converting to levels). This was done by running a nonlinear least squares regression based on the orthogonality condition  $\mathbb{E}[r_{ijt} - \exp(\beta z_{fjt})] = 0$ . This NLS regression would not be subject to the same Jensen's Inequality bias as a standard log-linear OLS. We then did a within-sample comparison of the predicted revenues from the NLS and found they performed substantially worse than the two-step OLS. As a result, the values we report for the remainder of the paper will be those coinciding with the two-step OLS described in this section.

## 5.2 Second Stage

Using the predicted values of potential revenue from the first stage regression,  $\hat{v}_{fjt}$ , we estimate the second stage using the system of moment inequalities laid out in PPHI. The estimation is founded upon a "revealed preferences" assumption. That is, whatever profits a firm receives from its

---

<sup>40</sup> We use the values they calculate for the country India. Note that Broda, Greenfield, and Weinstein provide their elasticities for 3-digit harmonized system codes, whereas our data are 5-digit ASICCC codes. We accounted for this by building a concordance from 3-digit ASICCC codes to 3-digit Harmonized System codes. In cases where there was an imperfect matching (such as when several different HS codes corresponding to one ASICCC code) we averaged the associated elasticities. There were a few cases in which certain elasticities were "substantially" different from other elasticities within their HS category (that is, differing by half an order of magnitude or more). In these cases, we matched 5-digit ASICCC codes to 3-digit HS codes, to ensure that these particular values were not misapplied to the wrong products within the data.



actions must be at least as large as the profits it could have earned from some counterfactual course of action in its original choice set. (This notion is formalized in Corollary 1).

This estimation method does not allow us to obtain point estimates on the variables of interest; however it does allow us to establish upper and lower bounds on those variables, by determining which values of the variables are consistent with the observed firm behavior, or in the absence of any such values, what values minimize the deviation from the moment inequalities.

The estimation proceeds in several phases. In the first phase, we select observations from the data that will help us identify particular coefficients in  $\theta$ , the set of variables to be estimated. In the second phase, we aggregate those observations into moments, which take the form of a set of linear inequalities. Estimation of the identified set then becomes equivalent to solving a linear programming problem using these moment inequalities as constraints.

### 5.2.1 Selecting Observations for Moments

As explained in section 4.3, we search for one-period deviations to derive inequalities based on the theoretical model described in the paper. Each of these inequalities becomes one “observation.” We then aggregate these observations into moments by averaging them, and it is these final aggregated moments that are used for the estimation of the parameter vector.

Equation (30) in Corollary 1 gives the expression for a single such observation. We can rewrite this equation as  $\mathbb{E}[\pi_{fdt}|J_{ft}] \geq 0$ , where the  $d$  denotes a deviation at period  $t$  from  $b_t$  to  $b'_t$ . Using Assumption 3, we can express this conditional inequality as an unconditional moment inequality:

$$\mathbb{M}_k = \mathbb{E}[g_k(Z_{ft})\pi_{fdt}] \geq 0 \tag{35}$$

where  $g_k(\cdot)$  is a positive-valued weighting function, and  $Z_{ft}$  is the set of values we require to be in the firm's information set in Assumption 3.  $k$  is an index for the particular moment inequality we are considering,  $k = 1, \dots, K$ .

Selecting observations for the moments is therefore equivalent to choosing the weight functions  $g_k$  to isolate one-period deviations that can be used to identify the parameters of interest. These  $g_k$  are allowed to depend on any information present in the firm's information set in period  $t$ .

The process of observation selection involves searching for patterns of firm behavior that would be informative for identifying one of the variables in our model. All of the variables we are estimating in the second stage relate to costs the firm has to pay (or an abatement of those costs). Thus, we will identify a variable by finding cases where the firm paid the costs associated with a variable, and then compare them to counterfactuals in the firm's choice set in which it could have avoided payment of the cost (in all or in part).

Consider the following example for the distance term,  $\zeta_1^{SC}$ , which appears in equation (22). This term represents the abatement of startup costs the firm receives for sharing common inputs with its destination product. The following table represents a hypothetical firm's choice of whether to produce a particular product  $j$  in periods 1 and 2. The "actual" row represents the observed production decision of the firm. The "counterfactual" row represents a possible alternative decision that was in the firm's choice set in period 2. (Because we are doing one-period deviations, period 2 is the only period in which the counterfactual behavior deviates from the actual behavior of the firm). A "1" in the table below signifies production of the given product, while a 0 signifies non-production.

		t =	1	2	3
Actual	j	0	1	0	
	j'	0	0	0	
Counterfactual	j	0	0	0	
	j'	0	1	0	

In the table above, the actual, observed behavior of the firm is production of product  $j$  in period 2, and non-production of  $j'$  in periods 1, 2, and 3. We consider the counterfactual where, in period 2, the firm chooses to produce  $j'$  instead of  $j$ .<sup>41</sup> In this example, the firm produces neither  $j$  nor  $j'$  in period 3.

By Corollary 1, the expected profits the firm receives from its actual behavior must be at least weakly greater than the profits from the counterfactual. This allows us to write the following inequality:

$$\begin{aligned} & \mathbb{E} \left[ v_{fj2} - \mu_0^{fc} - \epsilon_j^{fc} - \mu_0^{sc} + \zeta_1^{SC} \phi_{jb_1}^1 + \zeta_2^{SC} \phi_{jb_1}^2 + \zeta_3^{SC} \phi_{jb_1}^3 - \epsilon_{fj2}^{sc} \middle| J_{f2} \right] \\ & \geq \mathbb{E} \left[ v_{fj'2} - \mu_0^{fc} - \epsilon_{j'}^{fc} - \mu_0^{sc} + \zeta_1^{SC} \phi_{j'b_1}^1 + \zeta_2^{SC} \phi_{j'b_1}^2 + \zeta_3^{SC} \phi_{j'b_1}^3 - \epsilon_{fj'2}^{sc} \middle| J_{f2} \right] \end{aligned} \quad (36)$$

Which reduces to:

$$\begin{aligned} & \mathbb{E} \left[ (v_{fj2} - v_{fj'2}) + \zeta_1^{SC} (\phi_{jb_1}^1 - \phi_{j'b_1}^1) + \zeta_2^{SC} (\phi_{jb_1}^2 - \phi_{j'b_1}^2) + \zeta_3^{SC} (\phi_{jb_1}^3 - \phi_{j'b_1}^3) \right. \\ & \quad \left. - (\epsilon_j^{fc} - \epsilon_{j'}^{fc}) - (\epsilon_{fj2}^{sc} - \epsilon_{fj'2}^{sc}) \middle| J_{f2} \right] \geq 0 \end{aligned} \quad (37)$$

Thus, the  $\pi_{f dt}$  found in equation (35) is merely the left-hand side of equation (37). The above equation shows what a typical observation would look like for this particular pattern of firm behavior. If we needed to form the lower bound of  $\zeta_1^{SC}$ , we would select those observations for which  $(\phi_{jb_1}^1 - \phi_{j'b_1}^1) \geq 0$ . That is, those observations for which the proximity to the actual product chosen (along dimension 1) is greater than the proximity to the counterfactual product. To see why this is, consider the simplified scenario in which all the differenced terms in equation (37) are zero, except for

---

<sup>41</sup> Note there are many other potential counterfactuals that could be considered in this setting, each of which would give rise to different inequalities. We focus on this one merely to give an example of the method.

$(\phi_{jb_1}^1 - \phi_{j'b_1}^1)$  and  $(v_{fj_2} - v_{fj'_2})$ . Also, ignore the conditional expectation operator. We will discuss it momentarily. Then, equation (37) becomes:

$$(v_{fj_2} - v_{fj'_2}) + \zeta_1^{SC} (\phi_{jb_1}^1 - \phi_{j'b_1}^1) \geq 0 \quad (38)$$

Looking at it this way, it becomes clear why having  $(\phi_{jb_1}^1 - \phi_{j'b_1}^1) \geq 0$  is desirable for establishing a lower bound for  $\zeta_1^{SC}$ , since it allows us to write (38) as:

$$\zeta_1^{SC} \geq \frac{(v_{fj'_2} - v_{fj_2})}{(\phi_{jb_1}^1 - \phi_{j'b_1}^1)} \quad (39)$$

which is clearly a lower bound on  $\zeta_1^{SC}$ . However, if it had been that  $(\phi_{jb_1}^1 - \phi_{j'b_1}^1) \leq 0$ , we would have had to reverse the direction of the inequality when dividing by that term, and equation (39) would have represented an upper bound instead.

Of course, when we actually write the moments, we write them not in terms of ex-post realized values of the gross revenue terms, but rather in terms of the ex-ante expected values of those terms, conditional on the information the firm had available in the period in which it was making its decision. This is because our assumptions do not require the firms' decisions to be ex-post optimal, but only ex-ante optimal. Thus, the  $v_{fjt}$  terms in equations (38) and (39) represented expected gross profits.

We were able to express the lower bound for  $\zeta_1^{SC}$  in a very simple form by assuming that many of the other terms from equation (37) simply equated to zero. In practice, however, that will almost never be the case. What this means is that the bounds for  $\zeta_1^{SC}$  will depend on the bounds for many of the other variables in the model, and vice versa. This is not necessarily a crippling obstacle for our estimation, since in the moment inequalities method, all of the bounds are simultaneously determined.

However, what this does mean for our estimation is that wider bounds for one variable will translate into wider bounds for the other variables that depend on it.

The pattern of firm behavior we used as a demonstration above is useful for finding a bound on  $\zeta_1^{SC}$ , but is less informative about other terms within the firms' profit functions. For instance, both  $\mu_0^{SC}$  and  $\mu_0^{FC}$  cancel out in equation (36). This is useful for estimating  $\zeta_1^{SC}$ , since it allows us to attain simpler bounds on that coefficient and thus estimate it with less ambiguity. However, this means that particular pattern of behavior is useless for estimating  $\mu_0^{SC}$  and  $\mu_0^{FC}$ . We instead use different patterns for isolating these other variables.

Choosing such patterns for use in the moment inequalities framework is a bit of an art form, the goal being to generate observations in such a way as to get unneeded terms to cancel out in order to best isolate the coefficient of interest. Due to the similarity of our model to MSZ, many of the patterns we use mirror the ones found in their paper.

Table 25 shows explicitly which patterns were used to bound each coefficient. In selection of our patterns, we always conditioned on two periods: the period for which we are considering the counterfactual deviation, and one period prior. Those periods are indexed in the table by  $t=0$  and  $t=-1$  respectively. A "1" in the table represents production of the given product, while a "0" represents non-production. As explained earlier in the paper, firms are excluded if they are unobserved in any of the periods on which we are conditioning, or in the period following the counterfactual deviation.<sup>42</sup>

---

<sup>42</sup> We also perform a version of the estimation on large firms, since they are sampled with probability 1 in the ASI, thus eliminating ambiguity that may arise from firms entering and exiting the sample. The results are found in the appendix.

**Table 25**

Coefficient	Bound	Product	Actual		Counterfactual		Description of Counterfactual
			t = -1	t = 0	t = -1	t = 0	
$\mu_0^{fc}$	lower	j	1	0	1	1	Halt production of j
	upper	j	1	1	1	0	Produce j for one additional period
$\mu_0^{sc}$	lower	j	0	0	0	1	Produce j
	upper	j	0	1	0	0	Do not produce j
$\zeta^{sc}$ (all)	lower	j	0	1	0	0	Produce j' instead of j
		j'	0	0	0	1	
	upper	j	0	1	0	0	Same as lower bound
		j'	0	0	0	1	

As the reader might have guessed from the earlier discussion, although the patterns used for estimating the upper and lower bounds of the  $\zeta^{sc}$  terms are identical, we can identify which bound we are estimating by further conditioning on the sign of  $(\phi_j - \phi_{j'})$  along the given proximity dimension under consideration.

There is one further complication to consider. As we have already stated, we can only condition our selection of observations on data in the firm's information set during the period in which the counterfactual deviation is occurring. This means we can condition on any number of periods into the past, but not on any periods that occur after the deviation, since those were not observable to the firm at the time. This means there are actually four patterns of firm behavior that we must consider when estimating the bounds on the  $\zeta$ 's<sup>43</sup>:

		t =	1	2	3					
Actual	j	0	1	0						
	j'	0	0	0						
Counterfactual	j	0	0	0						
	j'	0	1	0						

					1	2	3					
		0	1	1								
		0	0	0								
		0	0	1								
		0	1	0								

					1	2	3					
		0	1	0								
		0	0	1								
		0	0	0								
		0	1	1								

					1	2	3					
		0	1	1								
		0	0	1								
		0	0	1								
		0	1	1								

<sup>43</sup> We are fleshing out this explanation for the bounding of the  $\zeta$ 's, but the principle we are describing (i.e. that we cannot condition on future periods) applies to the selection of observations for each of our coefficients.

Each of the observations for those patterns would give rise to a separate type of inequality. For instance, in the second pattern above, the firm would have to pay the static portion of the startup cost,  $\mu_0^{SC}$  twice in the counterfactual case, once for product  $j'$  in period 2, and then again for product  $j$  in period 3, whereas in the actual case, the firm only has to pay it once. This means that in addition to the other variables above,  $\mu_0^{SC}$  will also appear in the bounds for the  $\zeta$ 's, since it cannot be differenced out in the second and third firm behavior possibilities above.<sup>44</sup>

Note that these potential effects on firm profits in period 3 are not meant to imply that we use two-period deviations in our estimation. In each of the examples given above, the only difference in firm behavior between the actual and counterfactual cases occurs in period 2. Rather, we are saying that because firm profits are at least partially dependent on the state of the firm in previous periods, actions taken in period 2 can cause profits in period 3 to be different in the actual vs counterfactual cases, even if the period 3 actions of the firm are identical in both of those scenarios.

### 5.2.2 Aggregating Observations into Moments

After selecting observations in the manner described in the previous section, it remains to aggregate those observations into moments to be used in the estimation.<sup>45</sup> The theoretical moment inequalities are of the form given in equation (35). Thus, the sample moment inequalities are obtained by averaging all of the observations associated with a particular moment inequality, as follows:

---

<sup>44</sup> We do impose *one* restriction on the future in selecting our observations, and that is that the firm must actually be observed in all three periods of the search pattern. Because we need to know the firm behavior following the counterfactual period in order to fully compute the desired bound, if the firm does not appear in the dataset in the third period of our pattern, we drop that observation for being incomplete.

<sup>45</sup> A reader might wonder why we do this at all. If we have two observations, one saying  $x > 4$  and another saying  $x > 10$ , why not just say  $x > 10$  and be done with it? Econometrically, such a procedure would have undesirable properties (such as being vulnerable to measurement error), and might be compared to a linear regression performed on a single observation.

$$\mathbb{m}_k(\theta) = \frac{1}{D_k} \sum_{f=1}^F \sum_{t=1}^T \sum_{d=1}^{D_{it}} g_k(Z_{ft}) \hat{\pi}_{f dt}(\theta, \hat{\beta}) \quad (40)$$

Thus, for each moment inequality, (indexed by  $k$ ), we are summing over all firms ( $F$ ), all periods ( $T$ ), and all possible deviations consistent with the assumptions in our paper ( $D_{it}$ ).  $\hat{\pi}_{f dt}(\theta, \hat{\beta})$  is the predicted difference in profits between the actual and counterfactual firm actions, which depends on predicted values from the first stage regression (a function of  $\hat{\beta}$ ) and the parameter vector being estimated in the second stage,  $\theta$ .<sup>46</sup>  $D_k$  is the total number of observations used to compute the sample moment  $\mathbb{m}_k$ . Note that since the weighting function  $g_k(Z_{ft})$  can be zero for some values of  $Z_{ft}$ ,  $\mathbb{m}_k$  is computed with only a subset of the possible deviations.

### 5.2.3 Estimating the Bounds

After aggregating the observations, the estimation procedure involves solving a simple linear programming problem with the sample moment inequalities as constraints, as well as some “common sense” restrictions we place on our estimation. These additional restrictions are 1) Since each of the parameters we estimate is a cost, we require the acceptable values to be weakly positive, and 2) the value of the abatement of the startup cost due to proximity cannot exceed the startup cost itself (i.e.

$$\zeta_1^{SC} + \zeta_2^{SC} + \zeta_3^{SC} \leq \mu_0^{SC}).$$

More formally, let  $\Theta$  be the parameter space for  $\theta$ , and let  $\Theta_{\mathbb{m}}$  be the set of all values of  $\theta$  that satisfy the moment inequalities (as well as our additional restrictions, listed above). Thus,  $\Theta_{\mathbb{m}} = \{\theta \in \Theta: \mathbb{m}(\theta) \geq 0\}$ , where  $\mathbb{m}(\theta)$  represents the set of all  $K$  of the moment inequalities  $\mathbb{m}_k(\theta)$ .

---

<sup>46</sup> Note that although we do not index it,  $\theta = (\mu_0^{fc}, \mu_0^{sc}, \zeta_1^{sc}, \zeta_2^{sc}, \zeta_3^{sc})$  is allowed to vary across sectors (that is, across 1-digit ASIC categories).



Then, the maximum value along the first dimension of  $\theta$  is given by:

$$\bar{\theta}_1 = \left\{ \theta \in \Theta_m : \theta_1 = \arg \max_{\tilde{\theta} \in \Theta_m} \tilde{\theta}_1 \right\} \quad (41)$$

The definitions for the minimum and maximum values along other dimensions of the parameter vector are analogous.

#### 5.2.4 Properties of the Error Terms

One of the advantages of the PPHI moment inequalities framework is that it does not require us to assume a specific functional form for the error terms. There are, however, some restrictions that must be applied to ensure that our estimated set contains the true value of  $\theta$ . These restrictions are encompassed by the following assumption:

**Assumption 4:**<sup>47</sup> *The error terms are such that*

$$\mathbb{E}[g_k(Z_{ft})(\epsilon_{fdt}^v + \epsilon_{fdt}^{fc} + \epsilon_{fdt}^{sc})] \leq 0 \quad (42)$$

Recall that  $\epsilon_{fjt}^v$  is the approximation error of our gross profit prediction,  $\hat{v}_{fjt}$  from the first stage regression, and  $\epsilon_{fjt}^{fc}$  and  $\epsilon_{fjt}^{sc}$  are the error terms from the fixed and sunk costs, equations (21) and (22), respectively. The  $d$  subscript (as opposed to  $j$ ) on these error terms found in equation (42) merely shows that Assumption 4 imposes restrictions on the differences in the  $\epsilon$ 's between the actual and counterfactual cases, and not on the  $\epsilon_{fjt}$ 's themselves.

However, following MSZ, we can impose conditions on the  $\epsilon_{fjt}$ 's that are sufficient for the satisfaction of Assumption 4: 1) The first stage estimation procedure yields a consistent prediction for

---

<sup>47</sup> Note that Assumption 4 is analogous to Assumption 3 in PPHI. The additional requirement in PPHI's assumption is trivially satisfied in our model by the fact that weight function for firm  $f$ ,  $g_k(Z_{ft})$  does not depend on the choices of firms other than  $f$ .

the expected gross revenues, and 2)  $\mathbb{E}[\epsilon_{fjt}^{fc}, \epsilon_{fjt}^{sc} | J_{ft}] = 0$ . The latter restriction imposes that the firm does not have information on the fixed or sunk costs that is unknown to the econometrician.

### 5.2.5 Confidence Intervals

Confidence intervals for our parameter estimates follow the procedure outlined in PPHI, with the adjustment made in Holmes (2011) to account for correlation between observations arising from the same firm. We refer the reader to the cited papers for details on how these are computed.

## 6 Results

The main results are presented here, in Table 26. Using the moment inequalities method in PPHI, we do not get point estimates for any of our coefficients. Rather, we get upper and lower bounds on the potential values that those coefficients can take. As an example, of how to interpret this, observe that the static portion of fixed costs,  $\mu_0^{fc}$ , takes a maximum value of \$29,910 per product in industry 1 (Animals, vegetables, and forestry), and a minimum value of \$31,120 per product in industry 8 (railways, ships, and other transportation equipment), indicating that fixed costs are much greater in industry 8, as one might expect.

The values on the  $\zeta$  coefficients are telling for the importance of the different distance measures in each industry. To interpret the  $\zeta$ 's, remember that the proximity measures were all projected onto a 0 to 1 space, with a proximity of 0 representing products that are as far away as possible from the given firm along the chosen distance measure, and a proximity of 1 representing products that are "immediately adjacent" to the firm along the given dimension of distance. Therefore, products with a proximity of 1 to a firm along the first distance measure (input similarity) will receive the full benefit of the startup cost abatement for that measure. Products with a proximity of 0 will not receive any such

abatement (though it is possible that such products are close to the firm along another measure, receiving startup cost abatement from that alternate source).

For example, consider animals, vegetables and forestry. The coefficient on  $\zeta_1^{SC}$  has a maximum possible value of \$66,520. This means that if a potential destination product  $j$  had an inputs-similarity proximity of 1 to a firm in that industry (meaning, the cost share of the inputs for  $j$  exactly mirrored the existing cost shares of the firm in the period prior to introducing  $j$ ), that firm would receive a maximum of \$66,520 reduction in the startup costs associated with beginning production of that product. If none of the firms products shared any inputs with product  $j$  (and  $j$  was similarly far from the firm along the other two dimensions of distance), then the firm would have to pay the full startup cost to begin production of  $j$ , which our estimates show to be between \$5700 and \$109,210.

**Table 26 – Baseline Estimation**

	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>
Industry:	<b>Animal, Vegetable, Forestry</b>		<b>Ores, minerals, gas, electricity</b>		<b>Chemicals</b>		<b>Rubber, plastic, leather</b>	
$\mu_0^{fc}$	4.04	29.91	27.82	171.17	22.93	170.60	8.37	35.94
$\mu_0^{sc}$	5.70	109.21	26.41	598.02	56.45	670.82	28.35	164.29
$\zeta_1^{sc}$	0.00	66.52	0.00	318.94	0.00	273.82	0.00	62.26
$\zeta_2^{sc}$	0.00	109.21	0.00	598.02	0.00	670.82	0.00	164.29
$\zeta_3^{sc}$	0.00	36.18	0.00	190.08	0.00	203.24	0.00	43.75
Industry:	<b>Wood, cork, paper</b>		<b>Textiles</b>		<b>Metals, Machinery</b>		<b>Railways, ships, transport</b>	
$\mu_0^{fc}$	4.88	25.14	6.68	41.58	12.15	58.38	31.12	154.14
$\mu_0^{sc}$	9.49	99.01	6.46	191.79	36.51	260.71	104.23	700.00
$\zeta_1^{sc}$	0.00	50.41	0.00	77.76	0.00	87.36	0.00	234.34
$\zeta_2^{sc}$	0.00	99.01	0.00	191.79	0.00	260.71	0.00	700.00
$\zeta_3^{sc}$	0.00	30.41	0.00	49.02	0.00	64.79	0.00	170.27

*Notes: Values expressed in thousands of 1982 dollars. An exchange rate of 9 rupees per dollar was used for the conversion from rupees.*

Adding a product with a proximity of 0 to your firm would provide no abatement of the startup costs along the given distance measure. In our model, for proximities between 0 and 1, the benefit

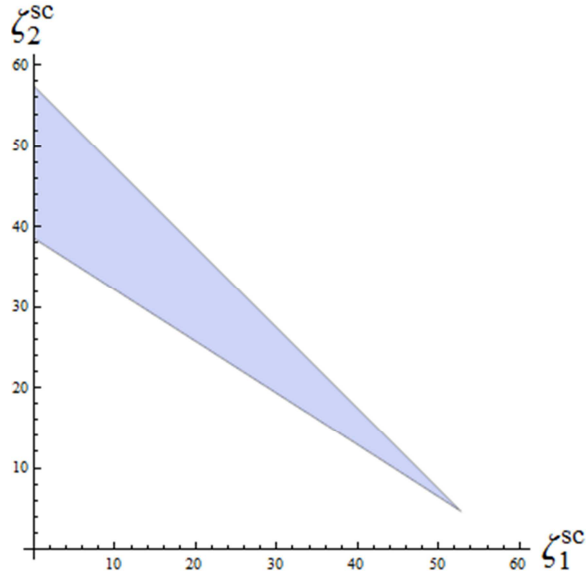
decreases linearly. So in animals, vegetables, and forestry, the maximum benefit of adding a product with a proximity of 0.5 along distance measure 1 would be  $\$66,520/2 = \$33,260$ .

It may appear from looking at the zeros in the table that it is possible that the distance measures do not matter at all. It should be noted, however, that the estimated set is *not* the Cartesian product of the upper and lower bounds presented in the table. Thus, just because the  $\zeta$  parameters all have 0 as their lower bound in the table, it does not follow that  $(\zeta_1^{sc}, \zeta_2^{sc}, \zeta_3^{sc}) = (0,0,0)$  is a point within the estimated set. Each one of the distance parameters might individually be zero, given certain choices for the other coefficients, but that does not imply they are jointly zero.

This is not easy to intuit just from looking at the table. The estimated set is a five-dimensional manifold, whose true shape is computationally difficult to determine, and even more difficult to represent in a two-dimensional picture. However, we can show a cross-section of the set, to illustrate to the reader that the bounds are not jointly zero. One such cross-section is presented in Figure 9.

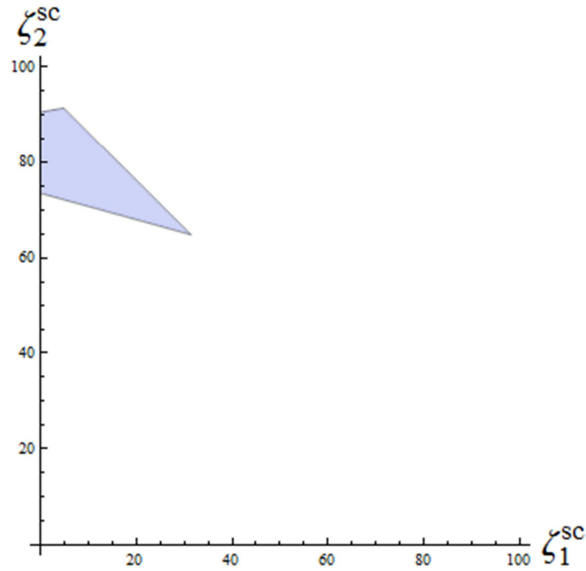
Figure 9 examines a cross-section of the estimated set for the Animals, Vegetables, and Forestry sector. We chose the median values of  $\mu_0^{sc}$  and  $\mu_0^{fc}$ , and  $\zeta_3^{sc} = 0$  to determine the location of the cross-section. We can observe from the picture that  $\zeta_2^{sc}$  is bounded away 0 for all values of  $\zeta_1^{sc}$ , and  $\zeta_1^{sc}$  is only 0 for particularly large values of  $\zeta_2^{sc}$ . Nevertheless, in this particular industry, at this particular cross section,  $\zeta_1^{sc}$  and  $\zeta_2^{sc}$  appear to be roughly substitutable.

**Figure 9 – Cross-Section of the Estimated Set for Animals, Vegetables, and Forestry**



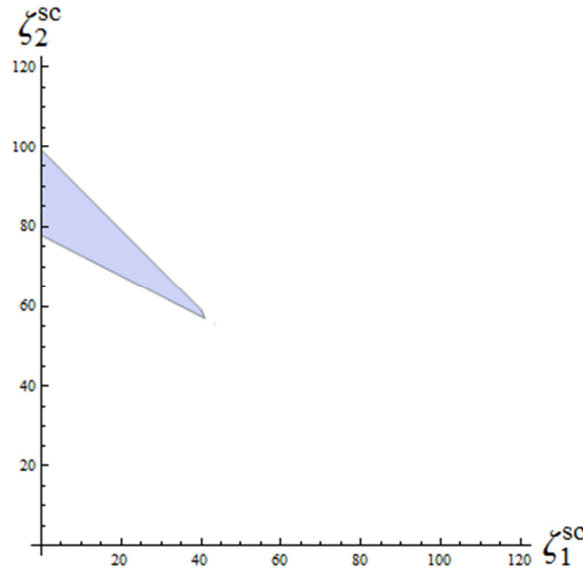
Notes: Values along the axes are thousands of 1982 dollars.  $\zeta_3^{SC} = 0$ , and the median values of  $\mu_0^{f^c}$  and  $\mu_0^{sc}$  were used to determine the position of the cross-section in the dimensions not shown in the picture.

**Figure 10 – Cross-Section of the Estimated Set for Rubber, Plastic, and Leather**



Notes: Values along the axes are thousands of 1982 dollars.  $\zeta_3^{SC} = 0$ , and the median values of  $\mu_0^{f^c}$  and  $\mu_0^{sc}$  were used to determine the position of the cross-section in the dimensions not shown in the picture.

**Figure 11 – Cross-Section of the Estimated Set for Textiles**



*Notes: Values along the axes are thousands of 1982 dollars.  $\zeta_3^{SC} = 0$ , and the median values of  $\mu_0^{f^c}$  and  $\mu_0^{SC}$  were used to determine the position of the cross-section in the dimensions not shown in the picture.*

The above characterization would not be true for every industry. For comparison, we present the cross sections for the Rubber, Plastic, and Leather Industry (Figure 10), as well as Textiles (Figure 11). In both of these industries, it is clearer that at the median values of the other variables, the range of potential values for  $\zeta_2^{SC}$  is much more limited. It is very clearly bounded away from zero by a significant amount in both of the figures.

The readers are referred to the appendix if they wish to see the linear inequalities that define the entire estimated set. Using these inequalities, it is possible to create cross-sections such as these for any choice of the other parameters in the estimation.

By examining the  $\zeta$ 's, we can receive some indication of which distance measures matter in which industries. In every industry, the ranking of relative importance for the three distance measures

seems to be the same. Merely looking at the maximum values, physical distance ( $\zeta_2$ ) seems to be the greatest contributor to product additions, followed by input similarity ( $\zeta_1$ ). The upstream/downstream connectedness measure ( $\zeta_3$ ) seems to fair the worst out of the three, consistently.

This is not to say that inputs and vertical connections are meaningless for product additions. Rather, that even at their maximum possible effectiveness, they tend to explain less of the variations in product additions than the physical distance component. On the other hand, there is a point in the estimated set for every industry in which the entire startup cost for new products in that industry can be abated by immediate physical proximity to the location of production.

Unfortunately, due to data limitations, it is not possible at this time for us to know precisely which portion of the production process is being helped by physical proximity. Many potential explanations come to mind, among them, knowledge sharing, access to natural resources, or local labor markets where workers have specialized skills. Distinguishing between these competing explanations is beyond the scope of the present paper, but we feel our results are a useful first pass, to indicate which areas of firm-product relatedness would be fruitful to investigate in the future.

Ninety-five percent single-sided confidence intervals for the baseline estimation are found in Table 27. While the estimated set specified by the confidence interval is obviously wider than that found in the estimation, the results are not dramatically different (with the exception of the chemical industry), ostensibly due to the large number of observations included in the estimation.

**Table 27 – Confidence Intervals for Baseline Estimation**

	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>	<b>Lower</b>	<b>Upper</b>
<b>Industry:</b>	<b>Animal, Vegetable, Forestry</b>		<b>Ores, minerals, gas, electricity</b>		<b>Chemicals</b>		<b>Rubber, plastic, leather</b>	
$\mu_0^{fc}$	4.04	35.53	27.82	207.50	22.93	221.16	8.37	42.00
$\mu_0^{sc}$	5.70	120.79	26.41	679.02	56.45	1,983.00	28.35	176.84
$\zeta_1^{sc}$	0.00	70.22	0.00	372.46	0.00	2,133.67	0.00	88.27
$\zeta_2^{sc}$	0.00	146.51	0.00	841.23	0.00	1,887.56	0.00	196.97
$\zeta_3^{sc}$	0.00	48.44	0.00	274.88	0.00	552.33	0.00	65.88
<b>Industry:</b>	<b>Wood, cork, paper</b>		<b>Textiles</b>		<b>Metals, Machinery</b>		<b>Railways, ships, transport</b>	
$\mu_0^{fc}$	4.88	32.49	6.68	49.68	12.15	68.60	31.12	184.60
$\mu_0^{sc}$	9.49	123.59	6.46	213.73	36.51	281.51	104.23	872.74
$\zeta_1^{sc}$	0.00	64.93	0.00	88.40	0.00	127.21	0.00	592.46
$\zeta_2^{sc}$	0.00	149.72	0.00	244.77	0.00	311.22	0.00	885.64
$\zeta_3^{sc}$	0.00	48.52	0.00	57.82	0.00	87.50	0.00	852.07

*Notes: Values expressed in thousands of 1982 dollars. An exchange rate of 9 rupees per dollar was used for the conversion from rupees. The left parameter in every column represents the single-sided 95% confidence interval on the lower bound, and the right parameter is the single-sided 95% confidence interval on the upper bound. Values account for correlation across observations, and were computed using 500 subsamples.*

## 7 Supplementary Analyses

To help us understand how the different channels affect firm behavior, we performed some calculations of potential firm product transitions using the model, and data from the estimation. Firms within this calculation determine profits in the way we have described in the theoretical model, with two notable exceptions: the degree to which firms are forward looking, and the calculation of the error terms.

In the model, we were not required to specify the degree to which firms are forward looking, because the moment inequality framework is consistent with a broad array of firm expectations and behaviors (see section 4.3). However, for the purposes of performing these calculations, this unbounded set of behaviors needs to be made finite and concrete. Our assumptions require that firms



take into account the effects of their current choice on static profit at least one period ahead. We therefore take this minimum required capacity for looking forward as the baseline for our calculation.

Secondly, within the PPHI moment inequalities framework, there are also relatively relaxed assumptions on the error terms (see section 5.2.4). However, for the purposes of our simulation, we draw the error terms from normal distributions with mean 0, which is consistent with the assumptions of the model. For the error terms associated with firm-product profits ( $\epsilon_{fjt}^v$ , see section 5.1), the standard deviation for the distribution is taken to be the actual standard deviation of a given firm's profits within its industry and year. For the other error terms ( $\epsilon_{fjt}^{fc}$  and  $\epsilon_{fjt}^{sc}$ , mentioned in 4.2.2 and 4.2.3, respectively), the standard deviation is taken to be  $\frac{1}{4}$  of the parameter estimate for the associated cost being used in the simulation.

The expected gross profits for each firm in the calculation are exactly the gross profit estimates we computed during our first-stage regression for the estimation. However, in order to mitigate the effects of some large outliers in the data, we dropped the top ten percent of the predicted profits. Firm locations are also identical to the actual locations found within the data.

We set the base year for the calculation to be 2000, and examined which products would be considered profitable by firms. For the second stage costs, we used the median values of the estimates from our baseline specification (those reported in Table 26). We excluded the upstream/downstream distance measure from the calculation due to its poor performance in the estimation.

This calculation, in addition to showing us the strength or weakness of our estimates also allows us to run counterfactuals, such as examining the results if we shut off or enhance one or both of the potential distance channels, or seeing the effect of the density of the firm-product connections on the number of profitable products.

## 7.1 Number of Profitable Products

For our first exercise, we examine the impact of negating the effect of each distance measure. Due to the amount of data produced by a calculation of this manner, we will only report one column of the output, in order to give the reader the basic intuition of how to interpret our results. Other rows within the output matrices follow the same general pattern.

The results of this exercise are reported in Table 28. Numbers in the table represent a count of the total products that have positive expected profits for firms whose main product is in ASICC category 21 (Salts, Sulpher, Lime, Cement). Stated another way, it is the sum of all the profitable firm-product relationships for firms in category 21. For example, imagine there are only two firms in category 21, A and B. Firm A has 3 potentially profitable products in Ores, and Firm B has 6 potentially profitable products in Ores. In that case, the entry in the table for Ores would be  $3+6 = 9$ . Thus, the table represents the number of possible expansion paths available to firms within that industry.

The first column of the table represents the result of these calculations for the baseline results. The second and third columns consider the counterfactual cases in which  $\zeta_1^{sc} = 0$  and  $\zeta_2^{sc} = 0$ , respectively. Setting  $\zeta_1^{sc} = 0$  effectively removes any benefit the firm might receive from sharing inputs with potential products. Similarly,  $\zeta_2^{sc} = 0$  removes any benefits it would receive from having production of a potential product located nearby.

**Table 28 – Profitable Products Available to Firms in Salts, Sulpher, Lime, and Cement**

	Baseline	$\zeta_1 = 0$	$\zeta_2 = 0$
Salts, sulpher, lime, cement	1750	1744	1146
Ores	110	110	70
Mineral fuels	391	391	264
Gas (fuel)	108	108	80
Electrical energy	154	154	107

Of note from the table is that negating the effect of the shared inputs does not substantially affect the number of profitable products at all, whereas negating the effects of local production affects it significantly.

Readers might be tempted to believe that this is an indictment against the shared inputs measure of similarity. However, it is necessary to interpret results within the context of the population distributions for the distances. In particular, observe the distribution for the input similarity measure. Most products are stacked up at 1. Products with a measure of 1 for this distance share no inputs with the firms' existing products, and thus receive no benefit from the cost abatement provided by  $\zeta_1^{sc}$ . Thus, setting  $\zeta_1^{sc} = 0$  does not affect the profitability for many products at all.

Alternatively, the distribution for the physical distances shows many products being produced in close proximity to the firm. These products will receive a substantial reduction in their startup costs from the physical proximity channel. Therefore, setting  $\zeta_2^{sc} = 0$  makes a big difference for a large number of products.

Therefore, the lesson to be learned from this exercise is that when interpreting the estimates, it is not enough to look only at the magnitude of the coefficients, but to consider also how those cost measures are interacting with the set of products in the firms' potential choice sets, and along which dimensions those products are "distanced" from the firm.

## 7.2 Firms' Product Choices

The previous exercise looked at all the profitable products available to the firm. In this exercise, we try to predict which products firms will move into, by allowing them to choose one product to add each period. For this simulation, we use the data for 2001-2002, since the earlier years of the sample were a little more sparse.

We'll motivate this exercise by showing the actual matrix of firm-product additions. The entries in the matrix show the number of firms that added a product in the column sector, conditional on having their main product in the row sector in the previous year.

**Table 29 – Actual Product Additions (Base Metals and Machinery)**

		Count of firms adding products in given sector								
Main sector in previous year		71	72	73	74	75	76	77	78	79
71	Iron, steel, & articles	283	11	35	47	28	19	13	5	9
72	Copper, nickel, zinc, & articles	2	21	7	6	0	1	1	0	0
73	Aluminum, tin, etc., & articles	14	4	47	13	3	3	8	0	1
74	Misc. manuf. Articles	58	10	14	35	39	23	17	5	7
75	General purpose mach. (non-elec)	57	3	7	26	155	69	56	7	9
76	Industry-specific mach. (non-elec)	34	2	4	16	67	158	31	6	15
77	Electrical machinery	43	17	30	26	63	35	259	51	16
78	Electronics equipment	6	2	2	3	7	5	34	82	1
79	Special purpose machines	12	5	3	6	13	12	15	2	21

Next, we will show the results from our simulation.

**Table 30 – Simulated Product Additions (Base Metals and Machinery)**

		Count of firms adding products in given sector								
Main sector in previous year		71	72	73	74	75	76	77	78	79
71	Iron, steel, & articles	87	1	14	61	237	156	274	36	26
72	Copper, nickel, zinc, & articles	9	0	3	4	19	12	23	4	2
73	Aluminum, tin, etc., & articles	18	0	3	19	39	24	61	3	7
74	Misc. manuf. Articles	34	1	7	16	73	65	125	17	14
75	General purpose mach. (non-elec)	51	0	5	39	143	83	163	19	18
76	Industry-specific mach. (non-elec)	37	1	17	25	89	73	117	19	7
77	Electrical machinery	82	4	11	44	199	124	279	38	19
78	Electronics equipment	19	0	4	16	42	34	67	7	5
79	Special purpose machines	8	0	1	6	22	18	34	1	6

Observing the tables, it is worth noting that although the simulation does not make perfect predictions, it performs better than one might expect for a model of its simplicity. It certainly appears to perform better than a fully random model, or an overly simplistic model in which firms only produce what they produced in the previous period (which would generate a matrix of zeros).

In some categories, the predictions of the simulation are actually very close to what we observe in the data. It predicts 279 electrical machinery firms will add products in their own sector, compared with 259 in the data. Its prediction of 143 general purpose machinery firms adding products within their sector is also close to the observed 155. Many other categories also closely match the data. In broad terms, it captures that there are few products being added in sectors 72, 73, and 74, and few products being by firms specializing in those sectors.

However, the simulation also highlights some weaknesses of the model. The most notable difference from the data seems to be the model's over-prediction of the number of products being added in the machinery sectors (75, 76, and 77), except in a few cases. This disparity seems most pronounced when examining firms in sector 71 (Iron, steel, and articles thereof).

That said, given the simplicity of the model, and the small number of parameters we estimated, one would not expect the model to perform perfectly. We used a very simple regression to determine potential revenues, coupled with a cost structure with only four parameters (recall we excluded the vertical connectedness measure,  $\zeta_3^{SC}$ , from the simulation). Furthermore, we applied a sweeping estimation technique generally to all firms in all industries.

With a process as complex and varied as the evolution of product scope, we cannot hope to fully capture all of the nuances of firms' decisions with one procedure. There are certainly many other factors that could be affecting their choices, and it seems natural to believe that our model would not be a good fit for every sector in every industry. Nevertheless, for some sectors, the model seems to

perform fairly well, producing predictions that are qualitatively and quantitatively similar to what we observe in the data.

### 7.3 Network Density Regression

As we mentioned in the introduction, one of the key results found by Hidalgo et al. (2007) was that the network of connections linking products together in terms of their relatedness is not evenly distributed. Rather, it is dense (meaning, with many close connections) in some areas, and very sparse in others. Therefore, countries (or in our context, firms) positioned in the dense part of the network are in a position to take advantage of many more cost abatement opportunities than those in the sparse part of the network.

In the work by Hidalgo et al. (2007), they presented a visual representation of the areas of these areas of density by providing a picture of their network linking products together. Our network is substantially more complicated to represent, because the connections we analyze are between firms and products, not between the products themselves. Therefore, we proxy for this density by measuring exactly how much abatement each firm receives from its position within the network for its sector.

Specifically, we compute for each firm and year, the normalized distance of the firm to each product within its sector along each of the dimensions in our study, and multiply this by the median of the  $\zeta^{sc}$  abatement parameter associated with that distance. Summing these figures together for all products gives the total number of dollars of potential startup cost abatement that the firm receives for that year. We call this number the “Network Density.”

We then regress the number of profitable products the firm has each year on: the network density just described, the firm-year fixed effect from the first stage regression (representing the firm’s

idiosyncratic productivity shock for that year), and the size of the firm's product basket in the given year. The results are given in Table 31.

Even controlling for the number of current products and the firm productivity, the network density is still highly significant (the t-statistic for that coefficient is 107). The seemingly small value of the coefficient should be interpreted in the light of the very large values of the network density measure.<sup>48</sup>

It should come as no surprise that the amount of cost abatement a firm receives is positively correlated with the number of potentially profitable products it has available. Rather, the purpose of this exercise was merely to highlight, in rather unsophisticated way, that different firms receive different benefits from their connections due to the density or sparsity of the network around them. This is to reiterate and expand upon the lesson of section 7.1, that the value and meaning of the coefficients found in this paper must be viewed within the context of the network of firm-product connections they interact with.

---

<sup>48</sup> The average firm received ₹585,000,000 of abatement, and even the least-benefited firm had over ₹2,400,000.

**Table 31 – Network Density Regression**

Regressor	Number of Profitable Products
Network Density	0.0052*** (0.00005)
Firm-Year Productivity	23.14*** (0.091)
Basket Size	2.264*** (0.134)
Constant	50.87*** (0.345)
Observations	136608
R <sup>2</sup>	0.405

*Notes: Heteroskedasticity-robust standard errors reported in parentheses. "Network Density" is measured in terms of ₹100,000s of startup-cost abatement within the firm's own sector only.*

*\*\*\* Significant at the 1% level*

## 8 Conclusion

We approached the question of how firm's product mixes evolve with the hypothesis that connections between firms and potential products were driving their decisions about which products to produce. We proposed several potential channels by which these connections might manifest, and tested their relative significance by observing the actual behavior of firms as they added new products and measuring the degree to which those products were connected to the firm along each of these dimensions. The model was estimated using moment inequalities, a novel econometric technique that allowed us to approach a large-scale choice problem of this nature in a computationally feasible manner.



The results speak strongly in favor of our hypothesis—that product connections matter, and are part of the driving force behind the observed co-production correlations between products. The success of the estimation also shows that history matters for firms' product choice, since each of the distance measures looked at connections between firms and products in the year prior to actual production. Finally, we were able to gain some insight into the nature of which connections matter most in which sectors—physical distance seems to matter the most, followed by input similarity. Vertical connectedness ranks as the least important measure of relatedness, in every industry.

There were, however, several drawbacks to our estimation. The first is that our estimates, based primarily on firms adding products within their own industries, are not easily generalizable to firms moving across industries. The second is that, due to data limitations and the constraints of our estimation method, we were unable to account for a lot of richness that is obviously a factor in firms' production decisions (such as the presence of specialized capital, credit constraints, or demand complementarities). Our model and estimation method also do not account for potential effects from cannibalism or credit constraints, which could be relevant in a developing country setting.

Nevertheless, the results we found should be an important first step in unraveling a very rich problem, and should prove useful to those seeking to understand how firms (and potentially by extension, countries) expand their product scope and migrate from one industry to another during their process of development.

## 9 References

- Almon, Clopper (2000). "Product-to-Product Tables via Product-Technology with No Negative Flows." *Economic Systems Research*. Vol. 12, Issue 1.
- Arthur, W. Brian (2000). *Increasing Returns and Path Dependence in the Economy*. The University of Michigan Press.
- Arthur, W. Brian (1989). "Competing Technologies, Increasing Returns, and Lock-in by Historical Events." *The Economic Journal*. 99, 116-131.
- Bernard, Andrew, J. Bradford Jensen, and Peter Schott (2006). "Survival of the best fit: Exposure to low-wage countries and the (uneven) growth of U.S. manufacturing plants." *Journal of International Economics*. 68, 219-237.
- Bernard, Andrew, Stephen Redding, and Peter Schott (2010). "Multiple-Product Firms and Product Switching." *American Economic Review*. 100:1, 70-97.
- Bohlin, Lars and Lars M. Widell (2006). "Estimation of commodity-by-commodity input-output matrices." *Economic Systems Research*. Vol 18, Issue 2.
- Broda, Christian, Joshua Greenfield, and David Weinstein (2006). "From Groundnuts to Globalization: A Structural Estimate of Trade and Growth." *NBER Working Paper No. 12512*.
- David, Paul (1985). "Clio and the Economics of QWERTY." *American Economic Review*. Vol. 75, No. 2.
- De Loecker, Jan (2011). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity." *Econometrica*. Vol 79, No. 5, 1407-1451.
- Dixit, Avinash and Joseph Stiglitz (1977). "Monopolistic Competition and Optimum Product Diversity." *American Economic Review*. Vol. 67, No. 3.
- Eckel, Carsten and J. Peter Neary (2010). "Multi-product firms and flexible manufacturing in the global economy." *The Review of Economic Studies*. Vol. 77, No. 1.
- Flagge, Matthew (2015a). "Predictive Power of Proximity on Firm- and State-level Production in India." Manuscript.
- Flagge, Matthew (2015b). "Foundations of Product Proximity." Manuscript.
- Foster, Lucia, John Haltiwanger, and Chad Syverson (2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*. Vol. 98, No. 1, pp. 394-425.
- Gollop, Frank and James Monahan (1991). "A Generalized Index of Diversification: Trends in U.S. Manufacturing." *The Review of Economics and Statistics*. Vol. 73, No. 2, 318-330.

Goldberg, Pinelopi, Amit Khandelwal, Nina Pavcnik, and Petia Topalova (2010). "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India." *Quarterly Journal of Economics*. 125 (4), 1727-1767.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg (2001). "The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools." NBER Working Paper No. 8498.

Harrison, Ann and Andres Rodriguez-Clare (2010). "Trade, Foreign Investment, and Industrial Policy for Developing Countries." *Handbook of Development Economics*. Vol. 5. Elsevier B.V.

Harrison, Ann, Leslie Martin, and Shanthi Nataraj (2013). "Learning versus Stealing: How important are market-share reallocations to India's Productivity Growth?" *World Bank Economic Review*. Vol. 27, Issue 2.

Hausmann, Ricardo (2014, July 26). "The Real Raw Material of Wealth." *Project Syndicate*.

Hidalgo, Klinger, Barabasi, and Hausmann (2007). "The Product Space Conditions the Development of Nations." *Science*. **317**, 482.

Hidalgo, Cesar and Ricardo Hausmann (2009). "The building blocks of economic complexity." *Proceedings of the National Academy of Sciences*, Vol. 106, No. 26.

Holmes, Thomas J. (2011). "The Diffusion of Wal-Mart and Economies of Density." *Econometrica*. Vol. 79, No. 1, 253-302.

Kugler, Maurice and Eric Verhoogen (2012). "Prices, Plant Size, and Product Quality." *Review of Economic Studies*. **79**, 307-339.

Leamer, Edward (1987). "Paths of Development in the Three-Factor, n-Good General Equilibrium Model." *Journal of Political Economy*. Vol. 95, No. 5.

Lederman, Daniel and William F. Maloney. (2012). *Does What you Export Matter? In search of empirical guidance for industrial policies*. The World Bank.

Marshall, Alfred (1920). *Principles of Economics*. 8<sup>th</sup> Edition. London: MacMillan and Co., Ltd.

Matsuyama, Kiminori (1992). "Agricultural Productivity, Comparative Advantage, and Economic Growth." *Journal of Economic Theory*. 58, 317-334.

Morales, Eduardo, Gloria Sheu, and Andrés Zahler (2014). "Extended Gravity." Mimeo.

Mowery, David, Joanne Oxley, and Brian Silverman (1998). "Technological overlap and interfirm cooperation: implications for the resource-based view of the firm." *Research Policy*. Vol. 27, Issue 5, 507-523.

Nelson, Richard and Sidney Winter (1982). *An Evolutionary Theory of Economic Change*. Cambridge, Massachusetts: Harvard University Press.

Pakes, Ariel, Jack Porter, Kate Ho, and Joy Ishii (Forthcoming). "Moment Inequalities and Their Application." *Econometrica*.

Penrose, Edith (1959). *The Theory of the Growth of the Firm*. Oxford: Basil Blackwell.

Redding, Stephen (1999). "Dynamic Comparative Advantage and the Welfare Effects of Trade." *Oxford Economic Papers*. 51, 15-39.

Redding, Stephen (2002). "Path Dependence, Endogenous Innovation, and Growth." *International Economic Review*. Vol. 43, No. 4.

Rodrik, Dani (2006). "What's So Special about China's Exports?" *China & World Economy*. Vol. 14, No. 5.

Sandberg, Lars (1969). "American Rings and English Mules: The Role of Economic Rationality." *Quarterly Journal of Economics*. Vol. 83, No. 1.

Santos Silva, J.M.C. and Silvana Tenreyro (2006), "The Log of Gravity," *The Review of Economics and Statistics*, 88(4): 641-658.

Schott, Peter K. (2008). "The Relative Sophistication of Chinese Exports." *Economic Policy*. Vol. 23, No. 53.

Sutton, John (1991). *Sunk Costs and Market Structure*. Cambridge, Massachusetts: The MIT Press.

Wang, Zhi and Shang-Jin Wei (2010). "What Accounts for the Rising Sophistication of China's Exports?" *China's Growing Role in World Trade*. University of Chicago Press.

## 10 Appendix

The appendix will include details on how we performed some of the calculations in the paper, as well as providing the results from alternative ways of estimating the model. Section A.1 will review how certain terms in the first stage regression were calculated, as well as providing the regression results. Section A.2 discusses our method for calculating the physical distance between firms and products. Section A.3 presents the results for some alternative specifications. Section A.4 presents the moments used in the preferred specification. Section A.5 gives the results from the Kolmogorov-Smirnov tests comparing the firm-choice and population distributions for firm-product distances discussed in section 6 of the paper.

### 10.1 Marginal Cost Regression

#### 10.1.1 $PL_j$

$PL_j$  is the price of a unit of labor in production of product  $j$ . Computation of this variable requires computing the labor costs for each firm, and using that to impute the labor costs of each product.

We began by calculating the labor inputs (in rupees) for every firm-year. Because we are interested in workers actually involved in the production process, we only included workers in the following categories in Block E of the ASI data:

1. Male workers employed directly
2. Female workers employed directly
3. Child workers employed directly
4. Workers employed through contractors

5. Supervisory and Managerial Staff
6. Other employees

These categories excludes unpaid family members/proprietor/coop. members. The total wage bill was calculated as the sum of the wages/salaries paid to employees in the included categories, excluding bonuses, contributions to Provident and other funds, and workman and staff welfare expenses.

To calculate the labor cost for a product, we need to make an assumption regarding how labor costs are assigned to given products within multi-product firms. We assumed that firms allocate labor expenses to products proportional to that product's share of the firm's total revenue from all products. So the labor costs allocated by firm  $f$  to product  $j$  in period  $t$  are:

$$Labor\ Costs_{fjt} = Labor\ Costs_{ft} * \frac{Revenue_{fjt}}{\sum_j Revenue_{fjt}} \quad (43)$$

We need to define what we will call a "unit" of labor for the purposes of our production function, so we can calculate the cost of such a unit. We use man-days as our unit of choice, and we use an analogous relationship to the one given in equation (43) to assign man-days to products within multi-product firms (that is, we assume man-days are proportional to revenue).

We then computed values for the price of labor (defined as labor costs divided by man-days) of each product on the firm-year level. The median of these firm-year specific labor intensities was then taken as the ultimate value for the product-level labor intensity:

$$PL_j = \text{median}_{ft}\{PL_{fjt}\} \quad (1)$$

### 10.1.2 $PIC_{fjt}$ (Intermediate Input Costs)

The calculation of the intermediate input costs for each firm-product-year combination requires several steps, which we will go through in turn. We first need to compute an input-output table for products at the 5-digit ASICC level<sup>49</sup>, we then use this table to assign inputs to outputs at the firm level. Finally, having the quantity of the given inputs assigned to each output, we find the cost of these inputs by multiplying the unit value of the input provided in the data.

### 10.1.3 Input-Output Table

There is a vast literature on the computation of input-output tables. As described in Bohlin and Widell (2006), an assumption needs to be made about technology in order for an input-output table to be identified. The two most common assumptions in the literature are the Product-Technology Assumption (PTA) and the Industry-Technology Assumption (ITA). The PTA assumes that production of a particular product requires the same inputs, regardless of which industry it is made in. The ITA assumes that, within an industry, the same input mix is used for every product produced by the industry.

Almon (2000) provides a discussion about the merits and weaknesses of both of these assumptions, as well as a demonstration of the types of input-output tables that would be produced as a result of each of them. As one might expect, the ITA fares very poorly, and Almon describes the tables produced by such an assumption to be “massive nonsense.”

We use the PTA for our input-output table, and generate it using the linear constraints in the technique developed in Bohlin and Widell (2006). This method was chosen because it allows the use of

---

<sup>49</sup> This is a greater level of disaggregation than is available from the Indian government.

the PTA while avoiding the problem of negative flows (i.e. negative inputs being used in some outputs), as well as allowing generalization to the use of rectangular “Make” and “Use” tables<sup>50</sup>.

We make use of the constraints in their minimization problem to harvest the usage coefficients that can be exactly identified from the data. So computing the input-output table comes down to solving the following set of linear constraints:

$$U_{uf} = \sum_{m \in \mathcal{M}} \alpha_{umf} M_{mf}$$

$$\alpha_{umf} \geq 0 \tag{2}$$

$$\alpha_{um} = \text{mean}_f(\alpha_{umf})$$

In the above equations,  $U_{uf}$  is the quantity of input  $u$  that is used by firm  $f$ .  $M_{mf}$  is the quantity of output  $m$  that is made by firm  $f$ .  $\alpha_{umf}$  is the usage coefficient, which is the number of units of the input good  $u$  needed to make one unit of the output good  $m$ .  $\alpha_{umf}$  is firm-specific. The average of those coefficients is  $\alpha_{um}$ , which becomes an element of the input-output table. The set  $\mathcal{M}$  is all of the products that the firm actually makes (in other words, we only apply the constraints for  $M_{mf} > 0$ ).

Intuitively, the outputs of a firm  $M_{mf}$ , times the quantity of input  $u$  that is needed to produce that output  $\alpha_{umf}$ , must equal the total amount of  $u$  that is used by the firm.

---

<sup>50</sup> The Make table is the mapping from producers to outputs. In our case, it is an  $F \times J$  matrix, where  $F$  is the total number of firms, and  $J$  is the total number of products. The element  $M_{fj}$  in the matrix gives the quantity of product  $j$  that was made by firm  $f$  in the given year (we have one Make table for each year). The Use table is analogous, but for inputs rather than outputs.



In the above equation, both  $U_{uf}$  and  $M_{mf}$  are known from the data, and we must determine  $\alpha_{umf}$ . We do this only for those  $\alpha_{umf}$ 's that are exactly identified from the constraints above. This happens in two cases.

In the first case,  $\mathcal{M}$  is a singleton, so the firm only makes one product. Thus,  $\alpha_{umf}$  is defined for every  $u$  for that firm and product (with  $\alpha_{umf} = 0$  for those products the firm does not use).

In the second case,  $U_{uf} = 0$  for some  $u$  and  $f$ . In that case, even if  $\mathcal{M}$  is *not* a singleton, we can determine that  $\alpha_{umf} = 0$  for that  $(u,f)$  because  $\alpha_{umf} \geq 0$  and  $M_{mf} > 0$ .

Intuitively, this method is roughly equivalent to using single-product firms to identify the elements of our input-output table, although the current methodology allows us to identify more elements of the table than merely using single-product firms.<sup>51</sup>

An input-output table was calculated using the above method for every year in the data. The final input-output table was then the median of the yearly tables.

#### 10.1.4 Assigning Inputs to Outputs at the firm level

Our estimation is performed on single- as well as multi-product firms, so we need a method to map a firm's inputs to its outputs in order to determine the input costs for a particular output.

---

<sup>51</sup> The above methodology allowed us to create a complete input profile for 3919 of our 5367 products, and a partial input profile for an additional 1099 of those products, leaving only 349 products for which no input data could be determined. Since many of our 5367 products only appear as inputs in the data (never outputs), this means we were able to calculate input data for almost all outputs in the dataset. With respect to the accuracy of this methodology, it is worth noting two points: 1) When computing the Gollop and Monahan (1991) distance measure between products, the distances looked qualitatively indistinguishable whether they were calculated using the input-output table above, or whether they were computed using firm input mixes (as in Kugler and Verhoogen [2012]), which incorporate multi-product firms and bypass the use of the input-output table (the formula for which is described in the "Theoretical Framework" section of the paper); and 2) The first-stage regression, which used intermediate inputs from the input-output table to predict marginal costs showed the coefficient on those inputs to be large and highly significant. Both of these facts lead us to conclude that this method, while not perfectly accurate, is at least a very good approximation to the "true" input-output matrix for these products.

Previous authors, such as Foster, Haltiwanger, and Syverson (2008) and DeLoecker (2011) address the problem of assigning inputs to outputs in multiproduct firms by assigning them in proportion to the number of products produced. We perform a similar operation, but unlike the aforementioned authors, we have the advantage of an input-output table which we can use to inform our assignment of inputs. We therefore modify their approach and weight the assignment of inputs according to the values found in the input-output table.

To do this, we assume there is a scaling factor  $\gamma$ , that relates firm-specific  $\alpha$ 's to the general economy-wide  $\alpha$ 's found in the input-output table, and that this scaling factor is constant for every product the firm uses. Consider the following illustration:

$\alpha$	$M_1$	$M_2$	$M_3$
$U_1$	.5	1	$\alpha_{13}$
$U_2$	$\alpha_{21}$	$\alpha_{22}$	$\alpha_{23}$
$U_3$	$\alpha_{31}$	$\alpha_{32}$	$\alpha_{33}$

Use	Quant
$U_1$	10
$U_2$	15
$U_3$	10

Make	Quant
$M_1$	5
$M_2$	10
$M_3$	0

The  $\alpha$  table is the economy-wide input-output table, in which we have only filled in two of the elements for this example, because we are only considering how to assign the input  $U_1$  to the firm's outputs. The Use table shows the quantity of each input used by our example firm, and the Make table shows the quantities of its outputs.

An *average* firm would need the following quantities of  $U_1$  to make the products of this example firm:

- $\underbrace{5}_{M_1} \times \underbrace{0.5}_{\alpha_{11}} = 2.5 = \text{amount of } U_1 \text{ needed to make 5 units of } M_1$

- $\underbrace{10}_{M_2} \times \underbrace{1}_{\alpha_{12}} = 10 = \text{amount of } U_1 \text{ needed to make 10 units of } M_2$

This firm would therefore need 12.5 units of  $U_1$  to make its existing set of outputs, but it only uses 10.

We therefore apply our scaling factor:

$$\gamma \underbrace{(M_1 \alpha_{11} + M_2 \alpha_{12})}_{12.5} = \underbrace{U_1}_{10}$$

In this example,  $\gamma = 0.8$ , so for the purposes of calculating the input costs for this firm, we would assume 2 units of  $U_1$  were used for  $M_1$ , and 8 units of  $U_1$  were used for  $M_2$ . When applying this method to the dataset,  $\gamma$  is allowed to vary by firm and use-product.

We use the above method to define a price for the total aggregated input basket used in production of each product at the firm-year level. Since most products in the data do not have units given in terms of quantity of items sold, we define a unit of output as being one rupee. We therefore divide the aggregated input costs for each product by the ex-factory value of output to determine the unit price for the input basket.

### 10.1.5 Regression Results

	$\ln(r_{fjt})$
$\beta_L^{mc}$	-0.779*** (0.039)
$\beta_{IC}^{mc}$	-0.130*** (.003)
<i>firm</i> × <i>year</i> FE	Yes
N	296677
$R^2$	0.75

\*\*\* denotes 1% significance.  
Robust standard errors are in  
parentheses.

Above are the results from the regression in equation (32), the first stage in our estimation procedure.

## 10.2 Physical Distance Calculation

### 10.2.1 Mapping firms to districts

There are two difficulties to be overcome in determining the location of the firms at the district level. The first is that the ASI panel data, which contains unique identifiers for firms, only gives firm location down to the state level, which is far less precise. Districts are available in the cross-section data, but there is no direct mapping from the cross-section to the panel. The second difficulty is that MOSPI changed their state and district codes in 2001. This required us to make two mappings: The first from the panel data to the cross-section, the second from pre-2001 district codes to post-2001 district codes.

To create the first mapping, from panel data to cross-section, we followed the technique used in Harrison et al. (2013), and matched the closing net value of fixed assets found in the panel and the

cross-section, dropping any values of 0 or 1, and any duplicates, which could potentially lead to ambiguous matches.

To create the second mapping, we made the assumption that firms (factories in the data), do not change their location from year to year. Thus, by observing the location codes of individual firms prior to and post-2001, we were able to create a concordance linking the two sets of codes.

### 10.2.2 Calculating the Distance

For each of the districts, longitude and latitude coordinates were obtained from Wikipedia's GeoHack tool. In the instances when coordinates were not available for a district, or when the available coordinates were obviously false, the coordinates for the district capital were used instead.

The coordinates were linked to the post-2001 district codes, because we did not have a list linking pre-2001 codes to district names. There were a few instances in which several pre-2001 codes were merged into one post-2001 code. In such cases, all of the pre-2001 codes were assigned the same coordinates.

Distances between the districts were calculated using the haversine formula for great circle distance, with the radius of the earth set to be 6372.8 km. Distances between firms were then defined to be the distance between the firms' associated districts, measured in kilometers, with a distance of 0 if the firms were located in the same district.

The distance between a firm and a product is then defined as the distance to the closest firm producing that product:

$$D_{fjb_{t-1}} = \min_{f' \in \mathcal{F}_{j,t-1}} d_{ff'} \quad (3)$$

Where  $d_{ff'}$  is the physical distance between firms  $f$  and  $f'$ ,  $D_{fjb_{t-1}}^2$  is the physical distance between firm  $f$  and product  $j$  at period  $t-1$ , and  $\mathcal{F}_{j,t-1}$  is the set of all firms producing  $j$  at  $t-1$ .

We then construct our measure of proximity by dividing by the maximum distance between any two points in India (to get the measure between 0 and 1), and flipping it, so that nearby products have a proximity measure of 1 instead of 0.

$$\phi_{fjb_{t-1}}^2 = \left| \frac{D_{fjb_{t-1}}}{\max_{f,f'} d_{ff'}} - 1 \right| \quad (4)$$

Where  $|\cdot|$  is the absolute value operator.

## 10.3 Alternative Specifications

### 10.3.1 Large Firms Only

Table 32 - Estimates

Industry:	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
	Animal, Vegetable, Forestry		Ores, minerals, gas, electricity		Chemicals		Rubber, plastic, leather	
$\mu_0^{fc}$	5.47	49.62	86.24	548.88	37.29	309.56	13.18	68.88
$\mu_0^{sc}$	7.12	173.50	74.73	1,857.67	74.05	1,200.11	40.81	304.72
$\zeta_1^{sc}$	0.00	111.06	0.00	862.52	0.00	489.51	0.00	115.00
$\zeta_2^{sc}$	0.00	173.50	0.00	1,857.67	0.00	1,200.11	0.00	304.72
$\zeta_3^{sc}$	0.00	62.86	0.00	602.68	0.00	362.92	0.00	79.11

Industry:	Wood, cork, paper		Textiles		Metals, Machinery		Railways, ships, transport	
	$\mu_0^{fc}$	13.60	85.32	8.99	58.60	21.55	118.90	55.45
$\mu_0^{sc}$	25.96	323.99	8.33	267.61	61.39	525.62	184.97	1,324.89
$\zeta_1^{sc}$	0.00	167.76	0.00	112.29	0.00	176.23	0.00	452.14
$\zeta_2^{sc}$	0.00	323.99	0.00	267.61	0.00	525.62	0.00	1,324.89
$\zeta_3^{sc}$	0.00	104.35	0.00	71.34	0.00	133.60	0.00	328.64

Notes: Values expressed in thousands of 1982 dollars. An exchange rate of 9 rupees per dollar was used for the conversion from rupees.

These are the results of our estimation performed only on the set of firms with 200 or more employees. According to the sampling procedure for the ASI, these firms are sampled with probability 1 in every year of the data.

Many of the broad trends identified in the baseline estimation persist. The physical distance parameter ( $\zeta_2^{sc}$ ) continues to have the largest upper bounds, followed by input similarity ( $\zeta_1^{sc}$ ), then vertical connectedness ( $\zeta_3^{sc}$ ). However, in this version of the estimation, both the lower bounds on the costs ( $\mu_0^{fc}$  and  $\mu_0^{sc}$ ) and the upper bounds on all parameters are substantially higher than in the baseline. This might be attributed to the larger scale operations happening at these firms, resulting in higher costs (but also potentially higher profits).

**Table 33 – Confidence Intervals**

Industry:	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
	Animal, Vegetable, Forestry		Ores, minerals, gas, electricity		Chemicals		Rubber, plastic, leather	
$\mu_0^{fc}$	5.47	62.72	86.24	673.19	37.29	417.22	13.18	81.70
$\mu_0^{sc}$	7.12	209.57	74.73	2,054.56	74.05	5,228.11	40.81	329.44
$\zeta_1^{sc}$	0.00	120.63	0.00	1,034.43	0.00	5,867.89	0.00	165.61
$\zeta_2^{sc}$	0.00	262.14	-0.03	2,602.22	0.00	4,802.56	0.00	373.71
$\zeta_3^{sc}$	0.00	87.86	0.00	1,019.66	0.00	1,884.56	0.00	119.79

Industry:	Wood, cork, paper		Textiles		Metals, Machinery		Railways, ships, transport	
	$\mu_0^{fc}$	13.60	117.18	8.99	70.99	21.55	142.27	55.45
$\mu_0^{sc}$	25.96	471.04	8.33	303.40	61.39	591.89	184.97	1,797.89
$\zeta_1^{sc}$	0.00	296.18	0.00	130.11	0.00	303.02	0.00	1,414.11
$\zeta_2^{sc}$	0.00	560.47	0.00	349.90	0.00	650.08	0.00	1,770.89
$\zeta_3^{sc}$	0.00	192.61	0.00	82.98	0.00	179.22	0.00	1,495.00

*Notes: Values expressed in thousands of 1982 dollars. An exchange rate of 9 rupees per dollar was used for the conversion from rupees. The left parameter in every column represents the single-sided 95% confidence interval on the lower bound, and the right parameter is the single-sided 95% confidence interval on the upper bound. Values account for correlation across observations, and were computed using 500 subsamples.*

The above table represents the confidence intervals for the specification including only firms with 200 or more employees. While for some sectors they are similar to the estimates themselves, in others (chemicals, ores, and transportation, for instance) they are much wider. This is likely attributed to fewer observations available in those sectors.



## 10.4 Moments for Baseline specification

**Table 34 - Moments**

		Industry 1						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.65	0.12	0.00	0.04	403,710	25,472
	lower	-1.00	0.21	-0.05	-0.01	-0.02	-28,299	8,363
$\mu_0^{sc}$	upper	1.00	0.66	-0.39	-0.60	-0.05	65,075	8,069
	lower	-1.00	-1.00	0.99	0.21	0.00	-20,657	40,598,000
$\zeta_1^{sc}$	upper	0.00	-0.28	-0.44	0.14	-0.05	-25,975	3,872,500
	lower	0.00	-0.17	0.20	0.20	-0.04	266,660	732,260
$\zeta_2^{sc}$	upper	0.00	-0.28	-0.42	0.09	-0.04	-20,092	3,369,100
	lower	0.00	-0.23	-0.12	0.34	-0.10	354,010	4,012,400
$\zeta_3^{sc}$	upper	0.00	-0.29	-0.31	0.19	-0.29	-30,693	700,130
	lower	0.00	-0.33	-0.10	0.29	0.21	322,360	123,730

		Industry 2						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.58	0.13	0.00	0.09	1,648,400	4,058
	lower	-1.00	0.19	-0.06	-0.01	-0.01	-218,380	2,017
$\mu_0^{sc}$	upper	1.00	0.66	-0.31	-0.59	-0.02	396,180	2,118
	lower	-1.00	-1.00	0.99	0.24	0.00	-47,464	12,068,000
$\zeta_1^{sc}$	upper	0.00	-0.28	-0.22	0.17	-0.04	-117,620	499,900
	lower	0.00	-0.18	0.20	0.21	-0.03	5,214,300	380,730
$\zeta_2^{sc}$	upper	0.00	-0.27	-0.16	0.09	-0.03	-95,104	467,680
	lower	0.00	-0.23	-0.05	0.37	-0.06	4,964,800	1,054,600
$\zeta_3^{sc}$	upper	0.00	-0.27	-0.13	0.24	-0.40	-95,989	62,616
	lower	0.00	-0.32	-0.12	0.23	0.13	2,181,400	23,879

		Industry 3						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.59	0.38	0.00	0.05	1,336,600	9,150
	lower	-1.00	0.19	-0.14	-0.01	-0.01	-162,720	3,706
$\mu_0^{sc}$	upper	1.00	0.73	-0.12	-0.64	-0.02	677,930	4,138
	lower	-1.00	-1.00	0.99	0.18	0.00	-80,319	17,261,000
$\zeta_1^{sc}$	upper	0.00	-0.27	0.09	0.21	-0.04	-154,710	1,417,700
	lower	0.00	-0.17	0.20	0.23	-0.02	960,350	998,920
$\zeta_2^{sc}$	upper	0.00	-0.25	0.11	0.11	-0.04	-128,230	1,138,300
	lower	0.00	-0.21	0.11	0.34	-0.04	1,706,500	3,653,900
$\zeta_3^{sc}$	upper	0.00	-0.28	0.11	0.21	-0.16	-164,500	400,830
	lower	0.00	-0.21	0.13	0.20	0.12	1,144,000	107,310

		Industry 4						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.54	0.33	0.00	0.04	594,480	4,491
	lower	-1.00	0.16	-0.10	-0.01	-0.01	-60,811	3,177
$\mu_0^{sc}$	upper	1.00	0.73	-0.18	-0.65	-0.02	246,970	3,291
	lower	-1.00	-1.00	0.99	0.16	0.00	-48,614	11,735,000
$\zeta_1^{sc}$	upper	0.00	-0.25	-0.03	0.18	-0.02	-70,731	1,059,100
	lower	0.00	-0.16	0.22	0.19	-0.01	324,880	648,950
$\zeta_2^{sc}$	upper	0.00	-0.23	0.00	0.09	-0.02	-57,363	869,600
	lower	0.00	-0.20	0.05	0.32	-0.02	467,740	2,042,400
$\zeta_3^{sc}$	upper	0.00	-0.28	-0.04	0.20	-0.16	-59,121	170,360
	lower	0.00	-0.23	0.12	0.22	0.16	409,280	43,694

		Industry 5						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.61	0.25	0.00	0.02	283,990	4,860
	lower	-1.00	0.18	-0.09	-0.01	-0.01	-36,280	2,204
$\mu_0^{sc}$	upper	1.00	0.68	-0.28	-0.60	-0.03	79,402	2,452
	lower	-1.00	-1.00	0.99	0.19	0.00	-29,549	11,500,000
$\zeta_1^{sc}$	upper	0.00	-0.31	-0.13	0.23	-0.04	-37,488	979,580
	lower	0.00	-0.16	0.20	0.21	-0.03	313,400	323,980
$\zeta_2^{sc}$	upper	0.00	-0.31	-0.09	0.15	-0.04	-29,973	757,730
	lower	0.00	-0.21	-0.09	0.33	-0.03	288,130	1,457,400
$\zeta_3^{sc}$	upper	0.00	-0.30	-0.12	0.23	-0.21	-27,443	198,710
	lower	0.00	-0.28	0.02	0.24	0.23	431,800	32,707

		Industry 6						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.52	0.18	0.00	0.06	429,470	10,946
	lower	-1.00	0.15	-0.06	0.00	-0.03	-54,524	8,661
$\mu_0^{sc}$	upper	1.00	0.73	-0.34	-0.69	-0.11	86,646	6,916
	lower	-1.00	-1.00	0.99	0.17	0.00	-23,197	29,306,000
$\zeta_1^{sc}$	upper	0.00	-0.26	-0.23	0.16	-0.16	-28,815	1,830,200
	lower	0.00	-0.15	0.20	0.18	-0.03	421,220	796,440
$\zeta_2^{sc}$	upper	0.00	-0.26	-0.19	0.11	-0.15	-24,490	1,642,900
	lower	0.00	-0.18	-0.08	0.29	-0.05	444,300	2,856,100
$\zeta_3^{sc}$	upper	0.00	-0.32	-0.30	0.21	-0.56	-14,972	509,780
	lower	0.00	-0.24	-0.07	0.19	0.19	432,050	143,780

		Industry 7						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.54	0.35	0.00	0.04	911,670	19,401
	lower	-1.00	0.15	-0.11	-0.01	-0.01	-91,434	15,278
$\mu_0^{sc}$	upper	1.00	0.75	-0.15	-0.68	-0.02	334,670	15,969
	lower	-1.00	-1.00	0.98	0.17	0.00	-54,246	48,288,000
$\zeta_1^{sc}$	upper	0.00	-0.24	0.02	0.15	-0.02	-85,650	4,957,200
	lower	0.00	-0.16	0.23	0.18	-0.01	649,930	3,291,400
$\zeta_2^{sc}$	upper	0.00	-0.23	0.04	0.08	-0.02	-74,199	4,245,200
	lower	0.00	-0.18	0.08	0.31	-0.02	674,990	9,073,800
$\zeta_3^{sc}$	upper	0.00	-0.22	-0.04	0.13	-0.14	-87,244	938,590
	lower	0.00	-0.22	0.10	0.21	0.15	740,030	439,510

		Industry 8						
	Bound	$\Delta\mu_0^{fc}$	$\Delta\mu_0^{sc}$	$\Delta\zeta_1^{sc}$	$\Delta\zeta_2^{sc}$	$\Delta\zeta_3^{sc}$	$\Delta v$	Obs.
$\mu_0^{fc}$	upper	1.00	-0.55	0.36	0.00	0.01	3,619,500	2,334
	lower	-1.00	0.15	-0.11	-0.01	0.00	-223,200	1,536
$\mu_0^{sc}$	upper	1.00	0.77	-0.15	-0.69	0.00	920,590	1,746
	lower	-1.00	-1.00	0.98	0.17	0.00	-151,860	5,819,300
$\zeta_1^{sc}$	upper	0.00	-0.26	0.04	0.18	-0.01	-248,480	452,100
	lower	0.00	-0.15	0.23	0.20	0.00	1,583,900	403,300
$\zeta_2^{sc}$	upper	0.00	-0.26	0.08	0.11	0.00	-229,150	403,350
	lower	0.00	-0.18	0.10	0.33	0.00	3,942,400	1,025,100
$\zeta_3^{sc}$	upper	0.00	-0.37	0.13	0.26	-0.05	-155,790	60,322
	lower	0.00	-0.22	0.09	0.22	0.07	2,137,500	67,471

Notes: Differences in profits are expressed in 1982 rupees. Besides the restrictions imposed above, we also impose the restrictions that the sum of the startup-cost-abatement parameters ( $\zeta$ ) cannot be larger than the total startup cost  $\mu_0^{sc}$ , and that no costs in the estimation can be negative.