Dynamic comparative advantages and product diversification

Andres Felipe Trejos Medina

University College London (UCL)

Thesis submitted for the degree of Ph.D. in economics

I, Andres Felipe Trejos Medina confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis explores empirically and theoretically the determinants of the process of product addition at the firm level. This process is relevant because it is the main driver of product switching, which has been found to be a key source of economic efficiency. This thesis consists of three chapters, which are sequentially connected, although each of them is a self-contained research paper. In chapter one I present new empirical evidence about the process of product addition at the firm level by using a detailed dataset of Colombian manufacturing firms from 1992 to 2017. My main finding is that the products that require an input mix that is more similar to the input mix currently used by a firm are more likely to be added by such firm in the future, even 3 years ahead, and that this correlation between the similarity in the input mix and the probability of product addition is larger for firms with more skilled labour. In Chapter 2 I present a model that rationalizes this main finding from chapter 1, among other findings. This model features key input-related characteristics that make some new products more profitable for firms than others. It focuses on the evolution of firm-input-specific productivities that determine what potential new products are chosen by the firms to be produced. This model yields specific propositions about the process of product addition at the firm level. Finally, in Chapter 3 I present empirical evidence that is mostly consistent with these propositions. For this, I use the same dataset as for Chapter 1. Very importantly, I perform an estimation of the structural firm-input-specific productivities in this last chapter, in addition to a causality analysis. For this latter I use an exogeneous reduction in tariffs carried out by the Colombian government in 2011.

Impact Statement

Economic growth is one of the key drivers of social progress in every nation and city of the world. As a consequence, growth policies are very common both at the national and subnational level. In general, they aim at promoting the general growth of the economies, and/or the growth of specific regions or economic sectors. For a targeted growth policy to be successful, it should target specific determinants of economic growth. These determinants should be ideally identified at a sufficiently specific level as to be well targeted by specific policies. This work contributes to this purpose by identifying key firm-specific determinants of the process of product addition. This is relevant because product addition is the main driver of product switching, which is the process in which a firm adds new products to its product mix and (possibly) drops other products from it. In turn, product switching is a key source of efficiency and growth at the firm level. If firms grow, the sectors, regions and countries that they belong to also grow. Despite this importance of the process of product addition, the research about its main determinants at the firm level has gained relevance only in very recent years (less than a decade). This work contributes to this research by identifying very granular firm-specific factors that allow firms to add more easily new products to their product mixes. Very importantly, these factors are not just firm-specific, but also firm-product-specific. In other words, this work does not only identify what factors are relevant for firms to add new products in general, but also what products are more likely to be added, given possible improvements in these factors.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Prof. Kalina Manova and Prof. Lars Nesheim, for their invaluable guidance and support. In addition, I am also grateful to the Economics Department at the University College London (UCL) for their teaching, funding and all the resources that they allowed me to have access to during this process. I am eternally grateful to Mr. Francisco Medina for his invaluable, unique and life-lasting support and encouragement.

Table of contents

Product addition: new evidence of a key phenomenon

Andres Trejos

June 19, 2025

Abstract

Product switching at the firm level is a key determinant of economic growth. One of its main dimensions is the addition of new products by firms (referred to here as *product addition*). This work uses a dataset of Colombian manufacturing firms from 1992 to 2017 to examine what are the characteristics of the new products added by firms to their product mix, and how such characteristics interact with characteristics of firms themselves. My main finding is that the potential new products that require an input mix that is more similar to the input mix currently used by a firm are more likely to be added by such firm in the future, even 3 years ahead. This positive correlation between the similarity in the input mix and the probability of product addition is larger for firms with more skilled labour. All this suggests the existence of persistent firm-input-specific economic factors that determine product addition. Such factors must be theoretically and empirically analyzed by future works, given their importance for growth policies. This is what chapters 2 and 3 of this thesis do.

1 Introduction

This paper is the first of three chapters that constitute my Ph.D. thesis. It presents new evidence about the phenomenon of product addition by Colombian manufacturing firms. The second chapter presents a theoretical model that rationalizes such findings. Finally, the third chapter evaluates the empirical validity of the predictions yielded by the model presented in the second chapter.

The addition of new (presumably more profitable) products by firms to their production mixes and/or their dropping of other (presumably less profitable) products when needed is a very important economic phenomenon. This process is referred to here as *product switching.* It has been found to be an important source of reallocation of resources within U.S. manufacturing firms towards their more efficient use (Bernard, Redding and Schott (2010)), as it allows firms to move their resources to more profitable activities (by switching products). Given this, the phenomenon of product switching contributes as much as the phenomenon of firm entry and exit to the evolution of U.S. aggregate manufacturing output (*ibid.*). It is also an important determinant of variability in the aggregate economic activity. Namely, firms drop more products and/or add less products in recessions (as their productivities fall), and this reduces the aggregate consumption because the number of varieties fall and the mark-ups increase (as there is less competition in the market for each product). These procyclical reductions in the aggregate consumption lead to reductions in the aggregate output, deepening the recessive episodes (Guo (2019)).

A key dimension of product switching at the firm level is the process of *product addition*, which happens when firms add new products to their product mixes. Product addition has been found to be more prevalent than product dropping in the U.S. (Broda and Weinstein (2010)), which means that the former drives the process of product switching to a larger extent than the latter. Product addition at the firm level generates positive spillovers for other firms' output (Ornaghi (2006)), as it may reduce the cost of inputs for other firms, and the newly produced products can be replicated by other firms. If a case of product addition by a firm implies also a case of product addition for a country as a whole (because the product is not only new for the firm but also for the country), it might be a source of economic growth, depending on how sophisticated the new product is. This because the addition of new more sophisticated products to the product mixes of countries has been found to be correlated with increases in their GDP per capita (Hausmann et al. (2009))[1].

Given this relevance of the process of product addition for reallocation of resources, productivity and economic growth, identifying its firm-level economic determinants is very important for analytical purposes and also for public policy. The identification of different possible determinants should ideally give rise to different policies, as good policies should target only the relevant determinants. I discuss the policy implications of the main findings of this thesis about the firm-level determinants of product addition in its third chapter. In spite of their importance, these determinants have not been sufficiently explored yet, and they started receiving sufficient attention just in recent years.

It was only very recently that some authors published empirical works that analyze the possible determinants of product addition at the firm level for a few countries. These works tackle key questions such as what are the characteristics of the new products added by firms and what these characteristics tell us about the underlying economic determinants of product addition. This paper contributes to this efforts by presenting a set of new empirical facts that shed light about the possible determinants of product addition by Colombian manufacturing firms during the last decades.

In this paper I explore what product-level characteristics make some potential new products more likely to be added by Colombian manufacturing firms in the future to their production mixes than others, and also the way in which such characteristics interact with some characteristics of firms (namely, with skilled labour) [2]. To do so, I construct a comprehensive dataset with detailed infor-

[1] Hausmann et al. (2009) define sophistication in such a way that a product is considered more sophisticated if it is produced by fewer countries, and if it produced by countries with more diversified products mixes.

[2] I also explored other dimensions of product switching different from product addition (such as dropping of products, number of produced products, number of years that a product

mation of Colombian manufacturing firms during the period 1992-2017. This dataset contains yearly information for every Colombian firm with more than 10 employees and/or sales above 120k U.S. dollars. It includes product-level information of prices and quantities for every firm, both for inputs and outputs, in addition to firm-level key variables such as capital stock, sales and labour. Very importantly, it is possible to trace firms over time in this dataset. Chapter 2 explains in detail the construction process of this dataset and its characteristics and advantages.

My main contribution in this paper is to establish four key empirical facts about the process of product addition. As their relevance depends on the extent to which they contribute to a better understanding of such process, I present a possible economic interpretation of the respective fact in each case. Very importantly, these interpretations are the basis for the theoretical model that I present in the second chapter of this thesis.

The first fact is that some products had substantially higher conditional probabilities of being added by Colombian manufacturing firms than others between 1992 and 2017. To conclude this, I define a feasible set of potential new products for each firm (see section 3 for details), and then I calculate the average probability of addition of each product across firms and years, conditional on the fact that the corresponding product belongs to the feasible set. I find that a high share of products has a low (¡0.2) probability of being added, whereas around 11% have a probability of being added above 80%. This difference indicates that some products are systematically more likely to be added by Colombian manufacturing firms.

I interpret this first finding as evidence that there must have been economic characteristics of Colombian manufacturing firms or specific circumstances and contexts that made such firms more prone to add some new products to their product mixes than others between 1992 and 2017.

My second contribution is to establish an important introductory fact about the input-related characteristics[3] of the new products added by the Colombian manufacturing firms to their product mixes between 1992 and 2017. Namely, I find that firms had in this period a higher probability of adding to their product mixes those new products whose production requires input mixes that are more similar to the input mixes they used before, just as Boehm et al. (2019) found for Indian firms. Formally, I find a positive and significant correlation between the similarity in terms of the input mix between a potential new product p and a firm f in a specific year and the probability that f adds p to its product mix one year ahead [4].

remains in the product mix of a firm and dispersion of the product-level shares of sales within a firm, among others). However, none interesting correlation was found between them and the characteristics of products and firms available in the dataset used for this paper.

[3] Here I use the word *inputs* to make reference to the set of physical materials used by a firm to produce its products by transforming them jointly into these latter. This set does *not* include the different types of physical capital, as these are not directly transformed into products.

[4] Hereinafter I will use just the phrase "more similar" to make reference to the fact that a product is more similar to a firm *in terms of its input mix*. Similarly, I will use just the

I interpret this second finding as evidence that the economic characteristics that cause different probabilities of addition for different products by the Colombian manufacturing firms must be related to their use of inputs, as it is the similarity in the input mix what seems to matter for product addition. Very importantly, I hypothesize that these characteristics must be firm-product-specific, as they yield different probabilities of addition for different products within the same firm. I conclude this after discarding other possible firm's characteristics and other possible reasons for product addition, such as changes in the prices of products.

My third and fourth findings are the most important contributions of this paper, given their novelty and relevance. They both deepen on the aforementioned correlation between similarity and the probability of product addition. The third finding is that the advantage of the more similar products in terms of the probability of addition is larger in firms with more non-production workers. Formally, I find that the interaction of the number of non-production workers of a firm f and the similarity between this firm and a product p in a year t is positively correlated with the probability that f adds p to its product mix one year ahead, even when I control for the similarity and the number of non-production workers themselves.

I interpret my third finding as evidence that the firm-product-specific characteristics that cause different probabilities of addition for different products within the same firm must interact with skilled labour (proxied here by the number of non-production workers) in such a way that having more skilled labour in a firm magnifies to a larger extent its probability of addition of those products for which it has *good* or *high* firm-product-specific characteristics. In short, firm-product-specific characteristics seems to interact in a complementary way with skilled labour.

Finally, my fourth finding is that the correlation between future product addition and the interaction of the current similarity and the current number of non-production workers remains positive and significant over time. Namely, it remains positive and significant even when "future" means three years ahead. Formally, I find that Colombian manufacturing firms add more often in $t+1$, $t+2$ and $t+3$ those potential new products that are more similar in t to their production in terms of the input mix. Very importantly, this persistent difference in probabilities in favor of more similar products is larger in firms that employ more non-production workers.

I interpret my fourth empirical finding as evidence that the firm-product-specific characteristics that cause different probabilities of addition for different products within the same firm must persist over time to some extent. In other words, they do not seem to disappear to a full extent from one period to other, as the correlation between current similarity and future product addition persists for several years.

word "similarity" to make reference to the similarity *in terms of input mix between a product p and a firm f.*

Summarizing, the sequence of my empirical contributions allows me to conclude in this paper that firm-product-specific economic factors *related to the use of inputs* may have determined to some extent the addition of potential new products by Colombian manufacturing firms between 1992 and 2017. Very importantly, such factors seem to persist over time to some extent. In addition, the extent to which these factors yield the addition of potential new products depends on the quantity of skilled labour used by a firm. Conversely, the size of the relation of this skilled labour with the probability of addition of the potential new products seems to depend on these firm-product-specific economic factors.

Given all these characteristics of the aforementioned firm-product-specific factors, I hypothesize in this paper that it can be the case that Colombian firms had in the period 1992-2017 firm-input-specific productivities that allowed them to add to their production mixes those *products* that required intensively those *inputs* in which such firm-input-specific productivities were high. For example, if a firm was good at using glass in a year t and it had never produced windows or books, it is natural to hypothesize that it was more prone to produce windows than books in the subsequent years (as windows are more intensive in glass than books). These interactions of firm-input-specific productivities with product-specific intensities constitute what I initially called *firm-product-specific factors*. Even though this possibility is not the *only* consistent with the facts presented here, it *is effectively* consistent with them. In addition, I hypothesize that these firm-input-specific productivities persist over time to some extent, and also that they interact in a complementary way with the quantity of skilled labour used by a firm.

The remainder of this paper contains four more sections. Section 2 explains in detail what are the strands of literature that this paper contributes to, and how it is related with the most important works in each of them. The third section describes the dataset in detail and explains how it is used for the purposes of this paper. Section 4 presents in detail all the empirical facts summarized in this introduction and my economic interpretations of them. Finally, Section 5 presents the conclusions of this work.

2 Contributions to the existing literature

This paper contributes to several strands of literature, which have analyzed the phenomenon of product addition from different perspectives and at different levels of aggregation. There has been a natural sequence of relevant research questions about this phenomenon in the last decades, and this paper contributes especially to the strand that tackled questions about the characteristics of the potential new products added by firms.

The research questions about product addition during the last decades can be summarized in three key questions, which have been sequentially explored by different authors in the same chronological order that they appear here: (i)

Are some products more likely to be added than others in different contexts? In other words, are there patterns of product addition? (ii) If this is the case, what products are more frequently added? In other words, what are the patterns of product addition? And finally (and most importantly in terms of economic analysis), (iii) what economic factors explain such patterns, if they exist?

This paper contributes to the strand of literature that tackles the question (i) above (about the possible existence of patterns in the phenomenon of product addition), as it corroborates and reinforces the conclusive previous evidence of the existence of patterns of product addition. I find here that the Colombian manufacturing firms were much more likely to add some products than others between 1992 and 2017. This finding contributes to confirm the validity of the previous findings that both countries (Hausmann and Hidalgo (2006)) and firms (Boehm et al. (2019), Bernard and Redding (2010), MacDonald (1985)) follow patterns in their processes of product addition (this is, that some new products are systematically more likely to be added to the product mix of countries and firms and countries).

To start, my first contribution is consistent with the conclusions of the papers that find patterns of product addition at the aggregate level. Hausman and Hidalgo (2006) found that countries add to their product mixes products that are more similar to the products they produced before. They define similarity between two products as their empirical probability of joint production, using data of country-level product mixes during almost 70 years.[5]. A similar fact is found for international trade flows. Feenstra and Rose (2000) found that there is on average a recurrent order in which countries tend to add new products to their baskets of exports to the United States.

Related with this, Eaton et al. (2007) found that there is a non-stochastic order in which Colombian exporting firms tend to enter to the different foreign markets. Although this possibility is not explored by them, this might happen to some extent because of patterns in the process of product addition by these firms. Namely, it is possible that different countries demand different products (given their preferences and levels of income), and Colombian firms enter to the markets of each country when the new products of the latter (determined by their patterns of product addition) coincide with the products demanded by the former.

My first contribution is also consistent with the conclusions of the papers that find patterns of product addition at the micro level. Boehm et al. (2019) found that some potential new products were more likely to be added by Indian manufacturing firms in recent decades than others. Although other papers are about the process of joint simultaneous production (instead of the process of

[5] Even though this fact seems to be more related to the second question about the pattern of product addition than to the first one about whether there is effectively a pattern, I associate it to this latter because what Hausman and Hidalgo (2006) really found is just the existence of the pattern that countries tend to add products that have been added often by other countries with similar production baskets, and not any fact related to the similarity between the potential products and such baskets in any dimension, as the term "similarity" used by them may suggest.

product addition), joint production is necessarily preceded by product addition in some order. Therefore, my first contribution also reinforces their validity to some extent. This is the case of Bernard, Redding and Schott (2010), who found empirical evidence against the hypothesis that the identity of each product produced by a U.S. manufacturing firm is independent of the identities of the other products it produces. Similarly, Bernard, Redding and Schott (2011) found that some pairs of products are more likely to be jointly produced within a U.S. manufacturing firm than others.

This work also contributes to the strand of literature that has tackled the question (ii) above. Once it became commonly accepted that both countries and firms follow patterns when they add new potential products to their production mixes or exports baskets, the next natural question was what those patterns are at the aggregate level and at the firm level. In other words, what are the characteristics of the new products added by firms and countries to their production mixes? Different works have found different answers to this question in different temporal and geographical contexts. My contribution to this strand is to present evidence that Colombian manufacturing firms are more likely to add in the future to their product mixes new products that require input mixes that are more similar to the ones they use in the present. Very importantly, I also present evidence that skilled labour augments the size of this correlation between similarity and product addition, and that this augmented correlation remains significant even when I analyze product addition three years ahead.

In this second strand of literature, my paper is very closely related to two specific papers. Those two works stand out within this strand because they examine in detail if some specific characteristics of products are correlated with some dimension of product selection, just as I do in this paper [6]. Firstly, Boehm et al. (2019) found for Indian firms that a product k that is not produced by a firm f in a year t is more likely to be added by f to its product mix in *t+1* if k is used intensively as an input by f in t, and also if the production of k requires an input mix that is similar to the input mix used by f in t. In a more recent work with data for U.S. manufacturing firms, Ding (2020) found that a firm's sales of a particular product k fall on average when there is a positive demand shock to another of its products (called j here), but its sales of k actually increase when this positive demand shock to j occurs if j and k are sufficiently similar to each other in terms of the combinations of knowledge inputs required for their production (called *knowledge input mix* hereinafter). Very importantly, he found that this complementarity in demand does not exist between pairs of products that are similar to each other in terms of the combination of other (non-knowledge) inputs required for their production.

The results from Boehm et al. (2019) and Ding (2020) are contradictory to some extent, as the former finds that similarity is positively correlated with a dimension of product selection (product addition), whereas the latter finds that similarity is *not* positively correlated with another dimension of product selection (sales growth), unless the analysis is restricted to knowledge inputs.

[6] The dimension of product selection explored by Boehm et al. (2019) and me is the addition of new products, whereas Ding (2020) uses the growth of sales of existing products

Although it is possible that a potential new product is chosen to be produced because of similarity but similarity is subsequently irrelevant for the growth of its sales, it is more reasonable to expect that both dimensions of product selection are related with similarity in the same direction [7]. My work sheds light on this discussion, as I find the same result as Boehm et al. (2019). However, I find that the correlation of product addition with similarity is stronger in firms with more skilled labour, which might be consistent with Ding's (2020) findings if the Colombian firms with more non-production workers use more knowledge inputs. I do not analyze here if this is the case.

My findings about the importance of the similarity in the input mix for product addition are also related to the work of Guo (2019). Even though she did not analyze the possible role of similarity as a determinant of the pattern of product addition, she found a relation between the *availability* of inputs and product switching (which is mainly driven by product addition). Namely, she found that the physical availability of inputs matters for the product switching, as it determines the capacity of the firms to produce different products. This finding is related to the ones presented in this paper because inputs play a key empirical role in both cases. In my work, their use determines product addition. In hers, their availability determines product switching (which is mainly driven by product addition). The third chapter of this thesis presents an analysis of causality that is closely related to Guo's (2019) work, in the sense that I explore there if another dimension of availability (different from phyisical availability) has an effect on product addition. Namely, I explore there if an exogenous generalized cheapening of inputs has an effect on product addition.

My work also complements the conclusions from MacDonald (1985), who summarized and built on findings from several industrial organizations researchers about the determinants of the selection of industries towards which U.S. manufacturing firms decide to move when they add new products to their product mixes. This work identifies the new industries selected by the firms and not new products as I do. Therefore, my work identifies more granular firms' decisions. However, my work is still complementary to his because they both analyze possible determinants of the direction of the process of product addition (as entering a new industry necessarily implies producing at least one new product). MacDonald (1985) concluded that U.S. manufacturing firms between 1963 and 1977 entered more often into new industries that were more similar to the industries in which they already produced in terms of R&D and marketing intensities, and also into vertically related industries (this is, into industries whose products they used intensively as inputs in previous years).

My work also complements and extends findings from other works of this second strand of literature that do not explore if some specific characteristics of products are correlated with product addition (as I do), but instead present suggestive evidence of the nature of such characteristics. This is the case of Bernard, Redding and Schott (2011), who found that firms tend to add products that are *similar* to those that they produce already. However, they do not

[7] As long as the role of similarity reflects capacities or skills of firms to use some specific inputs

define similarity, but only exemplify it. By using a formal empirical definition of similarity and exploring its correlation with product addition, I deepen on this suggestive evidence and I contribute to reinforce its validity. Although I use the same definition of similarity as Boehm et al. (2019), I present important new facts. Namely, I find that current similarity (measured as in Boehm et al. (2019)) remains positively correlated with future product addition even three years later, and that this correlation is larger the more skilled labour a firm has.

Finally, my work is related to a set of papers that do not strictly belong to this second strand of literature as they do not establish characteristics of the products added by firms, but that establish important characteristics of the *firms* that add more products on average to their product mixes or to their exports baskets[8]. Bernard, Redding and Schott (2010) found that product addition is more likely to be performed by larger firms (both in terms of output and employment) and by more productive firms (both in terms of output per worker and total factor productivity). On the other hand, Freund and Pierola (2016) found that atypically large firms (in terms of size) add many more products to their exports baskets than the rest of firms[9]. My work is related to these works because I also find a positive correlation between a particular characteristic of firms and product addition. Namely, I find that firms with more skilled labour have a higher probability of adding new products to their product mixes. Very importantly, I find that this feature makes them especially prone to add more similar products in terms of the input mix. The product-specific reach of this latter conclusion is exclusive of my work within this set of works.

This work also contributes to the strand of literature that tackles the question (iii) above. Namely, this work sheds light about possible novel economic factors behind the patterns of product addition. My main contribution to this strand is to present an economic interpretation of the new evidence presented here that emphasizes on the novel role of possible dynamic and persistent firm-input-specific productivities as key determinants of product addition. Moreover, I theorize that the evolution of these productivities depends on the skilled labour of a firm. To the best of my knowledge, input-related factors with these features of persistence and dependence on skilled labour have not been theorized before.

Several models with microeconomic foundations have theorized about the economic factors that determine the patterns of product addition or of related dimensions of product switching. Ding (2020) presents a model in which two attributes of inputs (mobility across products and scalability) and the intensity in their use for the production of different products determine how the production of some products changes as a response to shocks in the demand for others.

[8] The process of selection of the products that compose the exports basket of a firm may differ from the process of selection of the products that compose its product mix. Typically, exporting a product is harder than selling it in domestic markets. However, the profitability of both types of sales depends typically on one same feature of the firm (productivity), with different thresholds in each case. In this sense, both processes of selection of products may be correlated, and it makes sense to review and mention here those works that have identified characteristics of the firms that add more products to their exports baskets.

[9] Namely, they found that the contribution of top 5 firms (by sales) account on average for 85% of the cases of addition of new products to the exports basket of a country, using for this analysis a sample of 45 countries of different sizes from different regions.

Closely related to this, MacDonald (1985) theorizes that firms transfer their intangible capital among different activities to diversify their production. Boehm et al. (2019) propose a model in which firms decide simultaneously what inputs to become good at using (by making input-specific investments) and what products to produce.

Even though I do not present a formal model of product addition in this paper, my interpretation of the new findings presented here complements to some extent the theories proposed by Ding (2020), Boehm et al. (2019) and MacDonald (1985), as I theorize that economic factors with novel key features determine product addition to some extent. Namely, I interpret such findings as evidence of the existence of novel key dynamic economic factors (firm-input-specific productivities) that persist over time. In addition, I theorize about the key role of skilled labour for the evolution of these productivities. These features of my interpretation differ from key features of the theories proposed by Ding (2020), Boehm et al. (2019) and MacDonald (1985), as I explain below.

The key properties of inputs do not persist over time in the model proposed by Ding (2020) (as neither his empirical facts nor his model explore dynamic correlations between input-based variables and sales growth). On the other hand, MacDonald (1985) does not deepen on the properties of the intangible capital. As for Boehm et al. (2019), the evolution over time of their proficiencies in the use of different inputs is not modeled by them, and skilled labour does not play any role in their model. Even though my interpretation is not the *only* consistent with the new evidence presented here, it is consistent with it. The second chapter of this thesis deepens on this contribution, as I present there a formal model of product addition that builds on my interpretation of the new evidence presented in this paper.

3 Description of the data

This work uses a very detailed and comprehensive database for the Colombian manufacturing sector. It is often called EAM, which stands for its initials in Spanish (*"Encuesta Anual Manufacturera"*). I will use these initials throughout this document. The Administrative Department for Statistics of Colombia -DANE- started gathering information about the manufacturing firms of Colombia in 1950. Nowadays, it surveys each year every Colombian manufacturing firm with more than 10 employees and/or sales above an amount in Colombian pesos equivalent to 120k U.S. dollars in 2021, approximately.

The EAM contains several very important modules of information that played a key role in the process of finding the empirical facts that I will present in the next section of this paper. One of them is a module of products. This includes detailed information of every product sold by each manufacturing firm. This information includes the unitary price of each product produced each year by each firm, and also the quantity sold of each of them. It also includes both of these variables (unitary price and sold quantity) for exports. In addition, the EAM contains a module of inputs, which includes data on the purchased quantity and the unitary price of every input used by every firm in each year. This module also contains imported quantity and unitary price of imports for every imported input of each firm every year.

The EAM contains also several firm-level modules. One of them includes detailed information of the capital stock of each firm, disaggregated by type of capital. Namely, the EAM includes the estimated value in Colombian pesos of three types of capital stock for each year: machinery and equipment, office equipment (which includes information and communication technologies), and transport and storage equipment. The EAM does not only include the value of the stock owned by the firm of each of these three types of capital, but also the yearly depreciation of each of them, and the value of investment on each type of capital. My checks that the value of each capital stock each year was equal to the value of the previous year less the depreciation plus the investment were successful.

The EAM also contains firm-level information of employment. This module includes the number of workers by the type of work they perform. In some years this disaggregation was more detailed than in others. Namely, for some years it is possible to know the number of people in charge of sales, the number of managers, the number of administrative staff, the number of professional staff working in plant production[10] and the number of plant operators[11]. Unfortunately, the information of workers is less disaggregated for other years. As a consequence, I had to include all the workers in only two categories in every year, with the purpose of having consistency across years: production workers (which includes plant operators and professional staff working in plant production) and non-production workers (who are defined as all the workers in charge of developing and conducting the economic, financial and administrative policies of a firm, together with sales staff, administrative staff and distribution staff). Very importantly, the category of non-production workers includes the research and development staff, if these tasks are performed by the firm.

It would be ideal for this paper to distinguish between managers, R&D staff and other non-production workers, as each category might have a different effect on product addition. However, this is not possible, as the firms reported a unique number of people for this whole category in the EAM in several years. In the next section I will use this total number as a proxy of skilled labour. This could be controversial to some extent as it includes sales staff and distribution staff, together with managers and R&D staff. However, it is still true that this number increases when a firm employs more managers, high-level managerial professionals and R&D professionals, and that it is unlikely that a firm increases to a large extent the scale of its sales and distribution staff without having more managers and high-level managerial professionals. I will discuss this point in depth in the next section.

The EAM includes some cost modules that contain relevant information that

[10] This category includes the professionals (such as engineers), technicians and technologists who work directly in the physical production department of a firm.

[11] This category includes all the workers in charge of assembling, manufacturing, production and packaging activities

is not being used for this paper, but many researchers could be interested in using in the future. Namely, these costs modules include information of costs of energy and water, telecommunication services, taxes, interests and rent, among many other variables. In some years it includes complete modules about energy use, which intend to characterize the evolution of main energy sources of firms.

Very importantly for this research, each firm and plant is uniquely identified in the EAM with a numerical code that remains the same every year, as long as the firm is surveyed (which depends on the thresholds explained before). This allowed me to trace each firm over time and to construct a panel dataset. As firms can entry and exit this dataset temporally or permanently (depending on the thresholds), there might be biases in my results in this paper. This might happen if the reasons for entering or exiting were correlated with my relevant regressors. I will discuss in the next section how I solved the problems that might arise because of this possibility. However, it is worth to mention here that the turnover is not very high in the EAM. On average, 95% of the firms surveyed in a specific year were also surveyed one year before, and 94% of firms surveyed in a year are also surveyed one year ahead.

I use information of the EAM from 1992 to 2017, as information of years before 1992 has several problems and its organization would have delayed this research substantially, and 2017 was the last available year in the database when the bulk of this research was carried out. This period includes several economic cycles (caused to some extent by changes in prices of the commodities), a process of general trade liberalization and the signing and implementation of a free trade agreement with the United States.

There is a code in the EAM that uniquely identifies a product in the modules of products and inputs. Products and inputs are uniquely identified by a code of the International Standard Industrial Classification (ISIC) revision 2 adapted for Colombia from 1992 to 2000, by a code of the Central Product Classification (CPC) version 1.0 from 2001 to 2012, and by a code of the CPC version 2.0 from 2013 to 2017. In all cases, the EAM uses the most disaggregated levels of each classification (8 digits for the ISIC rev. 2 and 9 digits for the CPC).

The fact that the EAM uses always the most disaggragated available codes to identify products and inputs is very convenient for this research, as it ensures that a code truly identifies a specific product or input to the largest possible extent. However, this also prevents me from performing analysis that required tracing products or inputs across years with different classifications, as there do not exist correlative tables for these levels of disaggregation, but for much more aggregated categories of products. I could use such levels of aggregation to gain traceability, but I would have lost specificity in the definition of products, which I valued more. As a consequence, I can only trace products within the periods 1992-2000, 2001-2012 and 2013-2017, which fortunately are not extremely short.

The modules of products and inputs are crucial for this work. Unfortunately, they are not freely available, as their free use could violate some Colombian laws that protect the privacy of information for some firms that might be identified even though firms and plants are anonymised. A good example is Reficar, the largest oil refinery of Colombia [12]. Reficar is located in the city of Cartagena. It produces a high share of the total amount of gasoline and diesel used in the country. Even though it is impossible to identify Reficar by its legal ID (as this latter is different from the identifier that was assigned to it in the anonymization process), it is possible just to search for manufacturing plants in Cartagena in the business of refining, and there will be just one plant with sales as high as to correspond to Reficar. This would allow anyone to see sensible information such as unitary prices by product.

Because of the reasons explained in the last paragraph, the modules of products and inputs can only be accessed from an office located at DANE's main building. Unfortunately, it is not possible to guarantee that codes that are left running by external researchers run in nighttime and for several days, as interruptions use to happen in the main system. For this reason, I had to use a random sample of firms, as using all firms with all their modules of products and inputs in all years would have implied processing times that exceeded by far the time I was allowed to stay in their office. The pandemic of Covid-19 exacerbated this problem, as DANE's main building was closed for almost a year, and then reopened gradually at a very slow pace, with times as restrictive as only 8 hours per week for several months.

The total number of firms included in the EAM has been approximately 8.000 per year in the last two decades. As I am using data from 1992 to 2017, this means that my full firm-year-level database has in total approximately 208.000 observations (8.000 firms per year multiplied by 26 years). As each firm produces on average 3.5 products in a year and uses on average 25 inputs in a year, my firm-year-product-level database has in total approximately 728.000 observations, and my firm-year-input-level database has in total approximately 5.2 million observations. This latter number reflects the nature of the problem.

As I will explain in the next section, I will need to calculate similarities between a product and a firm in terms of the use of inputs for some of my empirical findings. This requires the calculation of many dot products of two vectors of use of inputs for each product-firm combination for each year, with each vector having as many elements as inputs are available in the economy. Had I used my complete database, such long dot products would have had to be calculated a number of times equivalent to the number of available firm-year combinations (208.000) times the average number of products in the feasible set of a firm, which I will explain in detail below (approximately 20). I tried this, and unfortunately it exceeded the capacity of the computers available at DANE's office.

In order to bypass this problem, I take a stratified random sample of 2 percent of all the available firms in my full dataset, this is, 160 firms. Strata are defined by an interaction of two dimensions: quartiles of sales of firms in their initial year and number of produced products by firms in their initial year. This procedure guarantees that the sample includes firms that had different sizes in sales and different product scopes when they were born (or started complying with the sales and/or employment requirements to be included in the EAM). The

[12] It has the capacity to refine 150.000 barrels of crude oil per day.

fact that each firm's probability of inclusion depends on the two aforementioned variables at their initial year and not on their averages across years prevents me from selecting a sample in which firms that started being small and grew both in sales and in product scope were over-represented. As a consequence, my sample yields a firm-year-level dataset with approximately 4.000 observations, a firm-year-product-level dataset with approximately 14.000 observations and a firm-year-input-level dataset with approximately 100.000 observations. The next section presents in detail the findings I found by using these datasets, and the procedures I followed in each case.

4 Empirical facts

This section presents key novel empirical facts about product addition and a possible comprehensive interpretation of them. I mainly show here that the potential new products that require an input mix more similar to that used by Colombian manufacturing firms in the present are more likely to be produced by these latter in the future (even three years ahead), and that this correlation is larger in firms with more skilled labour. I interpret this as suggestive evidence of the existence of firm-input-specific productivities that persist over time.

All the results of this section are explored and interpreted at the light of a particular set of three hypotheses that constitute the basis of the other two chapters of this thesis. To start, I hypothesize that each firm has specific levels of proficiency at using different inputs, and these levels determine what new products a firm can profitably add to its product mix and what products not, depending on the intensities of each product in the different inputs. If a potential new product is intensive in glass and a firm is very proficient at using glass, it is natural to expect that it is profitable for such firm to add this product to its product mix. I called these levels of proficiency "firm-input-specific productivities" in the introduction, and I will use that name hereinafter. I call this hypothesis "Hypothesis 1" hereinafter.

I also hypothesize that the firm-input-specific productivities persist over time to some extent. This hypothesis is reasonable, as it is reasonable to expect that once a firm becomes good at using a specific input, this ability does not vanish immediately. I call this hypothesis "Hypothesis 2" hereinafter.

In addition, I hypothesize that firms with more skilled labour can add more easily their potential new products to their product mixes. This hypothesis is sound, as skilled labour can be defined as the set of skilled and creative workers who are in charge of innovation, new ideas and knowledge-intensive activities [13]. As long as the capacity of a firm to add new products to its product mix depends to some extent on skilled labour defined as before (which is reasonable), the hypothesis above is theoretically sound. I call this hypothesis "Hypothesis 3" hereinafter.

[13] Please see Sun et al. (2020) for several definitions of human capital at the firm level, which is quantified here by the amount of skilled labour at the firm level.

Very importantly, Hypothesis 3 has a key addition. Namely, I also hypothesize that firm-input-specific productivities persist to a larger extent over time in firms with more skilled labour. In other words, I hypothesize that having more skilled labour does not only make a firm more prone to add new products, but also allows it to keep its firm-input-specific productivities over time to a larger extent. As long as the capacity of a firm to sustain its firm-input-specific productivities over time depends on the set of skilled and creative workers who are in charge of innovation, new ideas and knowledge-intensive activities, this addition to Hypothesis 3 is sound.

I present below five novel empirical facts about product addition, using data from the survey of Colombian manufacturing firms described in section 3 for this. In all the five cases I explain why the exploration of the corresponding fact is important at the light of the hypotheses presented above, and I interpret the results also at the light of such hypotheses.

Fact 1: The Colombian manufacturing firms are more likely to add some potential new products to their product mixes than others.

The first fact that needs to be established in order to find suggestive empirical evidence in favor of the hypotheses previously stated in this section is that some products are more likely to be added by the Colombian manufacturing firms to their product mixes than others. This because different potential new products are surely different in their intensities in the different inputs, and according to these hypotheses this should imply (*ceteris paribus*) that different potential new products have different probabilities of addition (unless the firm-input-specific productivities are exactly such that even though products differ in their input intensities, their probabilities of addition are identical or very similar, which is very unlikely).

In order to analyze if some potential new products have indeed a higher probability of being added by a firm to its product mix in the future, I start by providing an unequivocal definition of *potential new product* and *probability of addition*. Firstly, I define a potential new product for a firm f in a year t as a product that belongs to its feasible set but is not produced by it in this year. The feasible set of firm f does not change over time. It includes all the products that were ever jointly produced by any firm (including f itself) in any year with any product ever produced by f.

As for the probability of addition, it is defined as the empirical probability of a potential new product in t (as defined above) of being added in $t+1$. For its calculation, I start by defining firm-product-year-specific dummies that equal one if the respective product is a potential new product for the respective firm in the respective year and is added by it to its product mix one year ahead, and zero if the respective product is a potential new product for the respective firm in the respective year and is *not* added by it to its product mix one year ahead. Then I calculate the average of these dummies for each product across all firms and years. Please notice that this average only exists for products that were ever a potential new product for at least one firm in at least one year.

I used the module of products of the EAM to identify the products produced by each firm in each year, which allowed me to identify the feasible set for each firm, and therefore the potential new products of each firm in each year. I subsequently used the set of potential new products of each firm in each year t and the set of products actually produced by each firm in each immediately next year $t + 1$ to calculate the empirical average probabilities of addition as explained in the previous paragraph for all the products for which such probability exists [14]. Figure 1 shows the histogram of these probabilities of addition.

Figure 1 shows that conditional on being potential new products for firms, some products are much more likely to be added to the product mix of firms than others. As expected, there is concentration in low values, as many products are not added in most of cases[15]. Moving to higher probabilities, there is relatively similar concentration in different categories between 20% and 40%, and then less concentration from 40% to 80%. There is a minor concentration of mass (between 2% and 3% of products) around probabilities of addition of 50% and 65%. Then there is some mass concentration exactly at 80%. Namely, around 8% of products have a conditional probability as high as approximately 80% of being added to firms' product mix. In summary, around 50 % of the potential new products have a probability of addition lower than 50%, and then there are some concentrations of mass in the probabilities of addition higher than 50% that account for the other 50% of the potential new products.

Summarizing, some potential new products are more likely to be added to Colombian manufacturing firms' product mixes than others. This is consistent with the Hypothesis 1 stated in this section that product addition is governed by the interaction of the input intensities of the products and the firm-input-specific productivities. This because if this latter were actually the case, some products (those intensive in the inputs in which the Colombian manufacturing firms have higher firm-input-specific productivities) should be more likely to be added to firms' product mixes. However, just finding different probabilities of addition for different products is far from being sufficient to conclude that there is evidence in favor of Hypothesis 1. All I know so far is that some products have higher probabilities of addition, but I now I need to analyze what are the characteristics that seem to make some Colombian manufacturing products more likely to be added than others, and if such characteristics are consistent with Hypothesis 1. That is what I do in the next subsection.

Fact 2: Colombian manufacturing firms are more likely to produce in the near future new products that require input mixes that are more similar to the ones they currently use.

I found before that some potential new products are more likely to be added than others by Colombian manufacturing firms. Given my initial purpose of establishing evidence in favor of Hypothesis 1, the next natural question is if there are some product characteristics that make some manufacturing products more

[14] As explained before, the probability of addition is defined only for products that were ever a potential new product for at least one firm in at least one year

[15] Feasible sets have on average 20 products, whereas a firm produces on average 3.5 products a year.

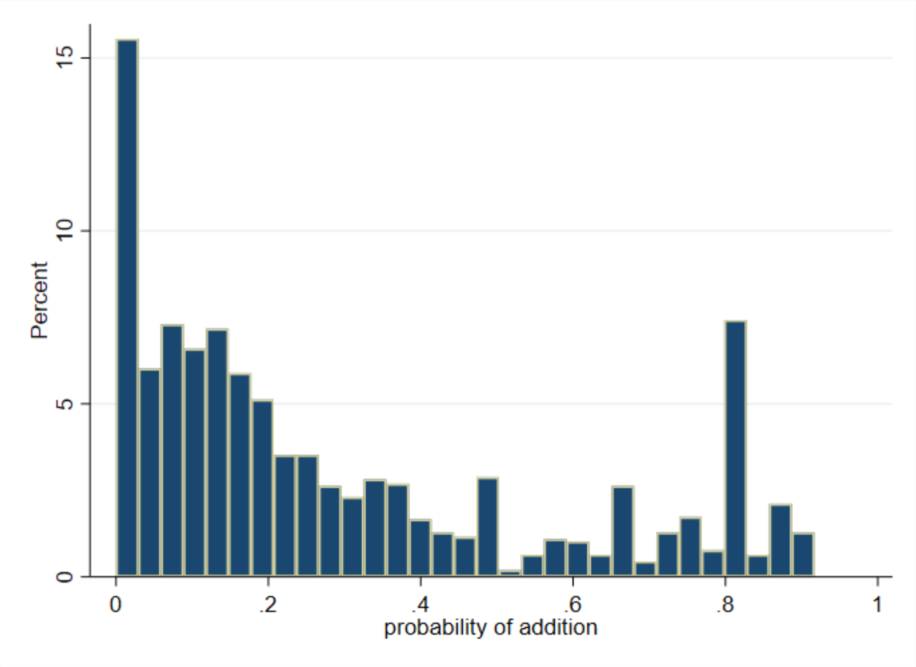


Figure 1: Conditional probability of addition

likely to be added by Colombian firms than others, and if such characteristics (if any) are consistent with Hypothesis 1. In order to answer this question, I identify different possible determinants of product addition according to previous studies and economic theory, and I analyze their possible correlations with the probability of product addition. For this, I estimate the parameters of a multivariate linear probability model in which the probability of addition of each firm's potential new products in each year depends on such determinants. I will emphasize on one of these possible determinants, given its closer relation to Hypothesis 1.

Firm-product-specific possible determinants

Previous literature provides several good candidates to analyze. Among them, there is one of essential importance for this work, given its closer relation to Hypothesis 1. Namely, Boehm et al. (2019) found for Indian manufacturing firms that the more similar is the input mix required to produce a new product to the input mix used by a firm, the higher is the probability of addition of this product by this firm one year ahead. This type of similarity (called just *similarity* here) [16] is crucial for my purpose of finding evidence in favor of Hypothesis 1, as if this hypothesis were true, then it would be precisely the potential new products with higher values of this *similarity* the ones with higher probabilities of addition. This because if each firm had input-specific productivities and these governed the process of product addition together with input intensities of products (as Hypothesis 1 states), then the new products produced by a firm

[16] See footnote 4.

in the future and the products it produces in the present should be similar in their input mixes (as both sets of products should be intensive in those inputs for which the firm has a high productivity). This expected similarity is precisely what the metric proposed by Boehm et al. (2019) measures. In summary, if Hypothesis 1 were true, the similarity proposed by Boehm et al. (2019) should be positively correlated with product addition.

Following Boehm et al. (2019), similarity is calculated as the dot product of the vector of a firm's expenditures in different inputs as shares of its total expenditure in inputs and a vector of the expenditures in different inputs needed to produce a product as shares of the total expenditure in inputs required for its production. Such product is normalized in such a way that it is always between zero and one. A zero indicates that both vectors are completely different, whereas a one indicates that they are identical. Formally, similarity between firm f and product p in t in terms of their input mixes is calculated as follows:

$$S_{fp}^{t} = \frac{\sum_{k=1}^{K} x_{fkt} x_{pkt}}{\left[\left(\sum_{k=1}^{K} x_{fkt}^{2} \right) \left(\sum_{k=1}^{K} x_{pkt}^{2} \right) \right]^{1/2}} \tag{1}$$

where x_{fkt} represents the share of total expenditure in inputs of firm f in year t that is spent in input k, x_{pkt} represents the share of total expenditure in inputs required to produce product p in year t that is spent in input k, and K is the total number of inputs available in the economy. This similarity is calculated for every possible combination of each firm with all its potential new products in each year (as defined in the previous subsection). Please notice that even though the feasible sets of firms do *not* change over time, the expenditure shares x_{fkt} and x_{pkt} *do* change over time, and therefore the similarity between a firm and the elements of its feasible set changes over time. The expenditure shares for a product p (this is, the set of x_{pkt} for different values of k) are calculated by taking the average of the expenditure shares of all the firms that produce the product p in year t. Finally, the expenditure shares for a firm f (this is, the set of x_{fkt} for different values of k) are directly observable from the module of inputs of the EAM.

Boehm et al. (2019) also found that the more a product is used as input by a firm in the present, the higher is its probability of addition one year ahead. This latter finding is consistent with the finding by MacDonald (1985) that U.S. manufacturing firms add more often products that belong to industries that are vertically related to their current production. Given these findings, I use the importance of the potential new product as an input in the past as another candidate to be correlated with the probability of product addition. Namely, I use the sum of the expenditures in the respective product in all the previous years of existence of the firm as an input divided by the total expenditure in inputs of the firm in all its previous years of existence. I calculate this variable for all the potential new products of all the firms in the sample in all their years of existence.

Firm-specific possible determinants

In addition to the variables explained before, I also use the firm's age as a candidate to be correlated with the probability of product addition. This because Klepper (1996) found that as firms become older, they devote more resources to process innovation relative to product addition. Similarly, Huergo and Jaumandreu (2002) found that entrant firms have higher average probability of innovation (including product addition), whereas such probability is lower for older firms. All these facts might be consequence of economic factors that could imply that older Colombian manufacturing firms add fewer new products to their product mixes.

If prices of all the inputs purchased by a firm fell from $t-1$ to t it would be reasonable to expect that it became easier for such firm to add all its potential new products, as the costs to produce all of them would fall. If only the prices of some inputs fell, the specific effect of this reduction on each firm would depend on how intensively the inputs with larger reductions are used by the firm. In order to account for this differential effect, I include the weighted average of the changes in the prices of all the inputs used by a firm between years $t-1$ and t as a regressor. For such average I use as weights the shares of each input in the total expenditure in inputs in $t-1$. It is reasonable to expect that the larger is this weighted average, the more difficult it becomes for the firm to add new products in general[17].

I also include the logarithm of the real value of sales of f in t as a regressor, in order to analyze if larger firms add more new products to their product mix on average.

Product-specific possible determinants

I also include as a regressor the average percentage change in the price of the potential new product between $t-1$ and t across all the firms that produced it in both years. This with the purpose of analyzing whether firms add more often products whose market prices grow more, as this might increase the profitability of producing them. I calculate this variable for all the potential new products of all the firms in the sample in all their years of existence.

For each potential new product of each firm, I also use as a regressor the share of each product's total past sales (until year t) by all the Colombian manufacturing firms that corresponded to exports. This share does not change by firm and product, but only by product. This variable attempts to capture the possible incentives of firms to add more often products that are sold to interna-

[17] Please notice that this intuition implicitly assumes that all the potential new products would require an input mix similar to the one used by the firm in $t-1$. This might not be the case. The new products might require an input mix very different to the one used by the firm in $t-1$, and therefore the inputs intensively required for the production of the new products might be more favored by the reduction in the prices of inputs very different to the inputs used by the firm in $t-1$. In order to explore this possibility, it would be necessary to quantify the differential effect of heterogeneous reductions in the prices of inputs on the cost of each potential new product of each firm, and then analyze if the firm-product-specific change in such cost affects the probability of addition. This is precisely what I do in an innovative causality analysis in the third chapter of this thesis.

tional (and possibly richer and larger) markets to a larger extent.

Linear probability model

In order to explore the role of all the previously mentioned regressors, I estimate the parameters of a multivariate linear probability model. I use as dependent variable a firm-product-year-specific dummy D_{fp}^{t+1} that equals one if the product p is added by the firm f in $t+1$, and zero otherwise. These dummies are defined only for the potential new products of each firm in each year. Therefore, the regression includes observations for all the potential new products of all the firms in the sample in each year. I use as regressors all the variables previously mentioned in this subsection. Formally, I estimate the parameters of the following model:

$$\mathrm{D}_{fp}^{t+1} = \alpha_0 + \alpha_1 S_{fp}^t + \alpha_2 age_f^t + \alpha_3 X_{fp}^{t,acc} + \alpha_4 EXP_p^{t,acc} +$$
$$\alpha_5 \Delta PriceProduct_p^{t-1,t} + \alpha_6 \Delta PriceMats_f^{t-1,t} + \alpha_7 logsales_f^t + \gamma_f + \gamma_p + \gamma^t + \epsilon_{fp}^{t+1}$$
(2)

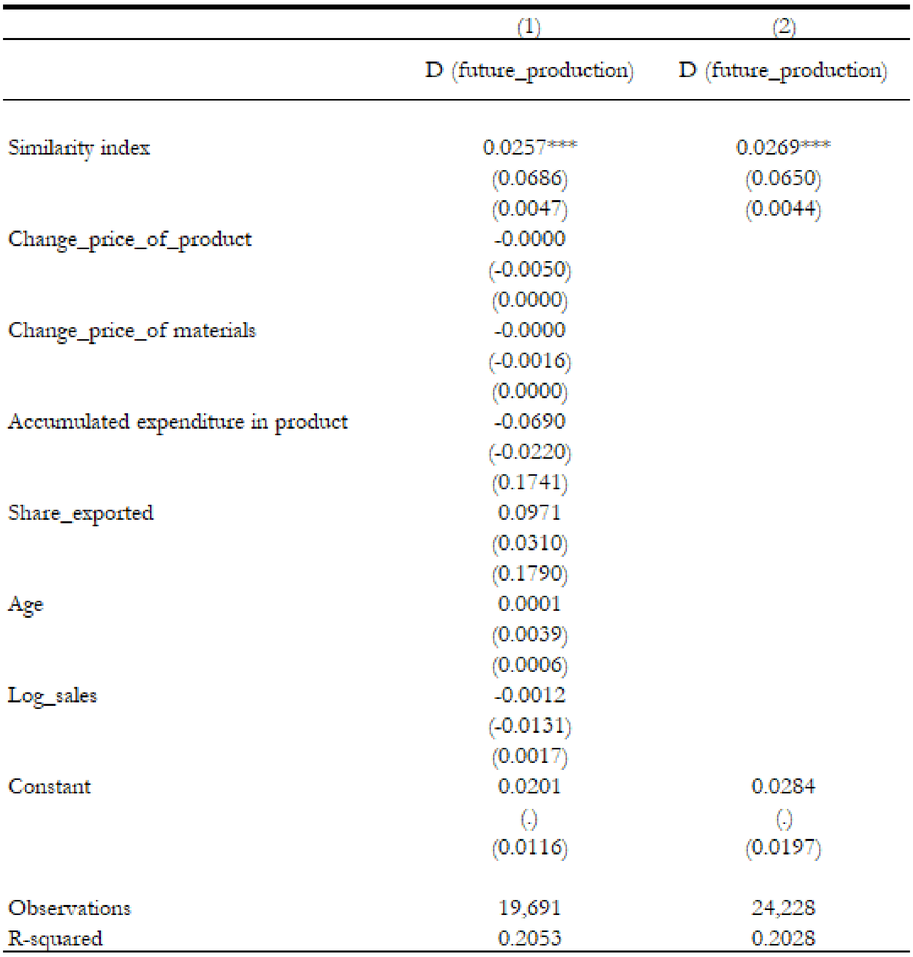
S_{fp}^t represents the similarity in the use of inputs between the firm f and the product p in year t, which is calculated by using the expression (1). On the other hand, $X_{fp}^{t,acc}$ represents the accumulated expenditure in product p as input until year t by firm f divided by the accumulated total expenditure in inputs of firm f until year t. $EXP_p^{t,acc}$ represents the share of total past sales of product p by Colombian manufacturing firms until t that were exported. $\Delta PriceProduct_p^{t-1,t}$ represents the average percentage change in the price of p between $t-1$ and t across all the firms that produced it in both years. Finally, $\Delta PriceMats_f^{t-1,t}$ represents the weighted average of percentage changes of the prices of the inputs used by f in $t-1$ and t, using shares of total expenditure in inputs in $t-1$ as weights.

γ_f, γ_p and γ^t are fixed effects by firm, product and year, respectively. I include them in order to capture otherwise omitted factors that might be correlated with the probability of product addition to some extent and that change only across firms, across products and over time, respectively.

Results

Table 1 presents the results of the estimation of the parameters in expression (2) by least squares with clustered errors by firm. The numbers in parentheses one line below the estimators are the beta coefficients. Once multiplied by 100, they quantify in this case the change in percentage points in the probability of addition of a potential new product associated with a change of one standard deviation in the respective regressor. The numbers in parentheses two lines below the estimators are the corresponding estimated standard errors. The Column 1 shows the results when all the regressors are included.

The similarity index is highly significant and has the expected sign. Namely, the probability of addition by Colombian firms one year ahead is higher for potential new products that currently require input mixes that are more similar to the input mixes used by the firms.



	(1)	(2)
	D (future_production)	D (future_production)
Similarity index	0.0257***	0.0269***
	(0.0686)	(0.0650)
	(0.0047)	(0.0044)
Change_price_of_product	-0.0000	
	(-0.0050)	
	(0.0000)	
Change_price_of materials	-0.0000	
	(-0.0016)	
	(0.0000)	
Accumulated expenditure in product	-0.0690	
	(-0.0220)	
	(0.1741)	
Share_exported	0.0971	
	(0.0310)	
	(0.1790)	
Age	0.0001	
	(0.0039)	
	(0.0006)	
Log_sales	-0.0012	
	(-0.0131)	
	(0.0017)	
Constant	0.0201	0.0284
	(.)	(.)
	(0.0116)	(0.0197)
Observations	19,691	24,228
R-squared	0.2053	0.2028

Table 1. Regressions of probability of addition of potential new products in $t + 1$ as a function of the similarity in the use of inputs and other regressors. Beta coefficients in parentheses one line below the estimators. Estimated standard errors in parentheses two lines below the estimators.

None of the other regressors included in the model of Column 1 is statistically significant even at a level of significance of 10%. This means that the variables that were found to be correlated with the probability of product addition in other countries in the past (incentives for vertical integration and firm's stage of evolution) do not seem to be correlated with the probability of product addition by the Colombian manufacturing firms. The same is true for two characteristics of the potential new products, namely their past export shares and the recent change in their market prices. None of them is statistically significant. Finally, I found that the average changes in the prices of inputs are not significantly correlated with the probability of product addition by the Colombian manufacturing firms.

Summarizing, just the measure of similarity is statistically significant. It is significant even for a level of significance as low as 1% and it has a positive sign, as expected. I excluded all the other regressors in column 2, in order to explore possible changes in the significance of similarity caused by possible high collinearity between this variable and the other regressors (although what should happen is that significance increases from Column 1 to Column 2 if such collinearity existed indeed). The parameter is found to be statistically different from zero in both cases even at a level of significance of 1%. In addition, I found that the size of the estimator does not change much from Column 1 to Column 2, which indicates that the exclusion of the other regressors in Column 2 does not seem to introduce a large bias to the estimator of α_1.

The beta coefficients shown in Table 1 indicate that an increase in the similarity of the size of one standard deviation is associated with an increase of approximately 6.5 percentage points in the probability of addition. This relation is not small in economic terms, although it is not very big either. This might reflect the fact that Colombian manufacturing firms do not add many products in general. As said before, the average firm produces just 3.5 products each year, and once a product is added to the product mix of a firm, its probability of continued production in the next year is very high (above 90% on average). These two facts necessarily imply that Colombian manufacturing firms do not add many products on average.

The finding that similarity in the input mix is statistically significant and positively correlated with the probability of addition is crucial for the possible validity of Hypothesis 1. This because this finding means in a strict sense that the Colombian manufacturing firms choose their new products in such a way that they are able to use intensively for their production the same inputs that they use intensively in the present. In turn, this latter suggests that there might exist heterogeneous firm-input-specific productivities, which might imply to some extent both an intensive use in the present of those inputs for which the firm has a high specific productivity, and a higher potential profitability of adding one year ahead those new products that are also intensive in those inputs. This is what I conjectured in Hypothesis 1. The fact that the addition of these more similar new products (in terms of their input mix) occurs one year ahead is also indicative of the possible validity of Hypothesis 2, as it indicates that the aforementioned firm-input-specific productivities in a particular year might remain to some extent one year ahead.

So far I have found that the probability of addition of potential new products by Colombian manufacturing firms one year ahead is higher for some products than for others, and that it is higher for those products that require for their production input mixes that are more similar to the input mixes used by the firms. These facts constitute suggestive evidence in favor of the Hypotheses 1 and 2. As I want to analyze if there exists possible evidence also in favor of Hypothesis 3, I proceed now to analyze the possible role of skilled labour in the relationship between similarity and the probability of product addition.

In my Hypothesis 3 I hypothesize that firm-input-specific productivities persist to a larger extent over time in firms with more skilled labour. If this were the case, the relation between the probability of product addition one year ahead and the similarity in the input mix in the present should be even tighter in these firms. This because the firm-input-specific productivities would remain more similar from the present to one year ahead in those firms, and I hypothesize (in Hypothesis 1) that these productivities determine what products are added by a firm and what products not. If this determinant of product addition is more similar in both periods in these firms, the produced products in both periods should be also more similar to each other in those firms in terms of the attribute that makes them more appealing for a firm at the light of its input-specific productivities, this is, their input mix.

Fact 3: The relationship between similarity in the input mix and future product addition is stronger in firms with more skilled labour.

In order to analyze if the relation between the probability of product addition one year ahead and similarity in the input mix in the present is indeed tighter in firms with more skilled labour, I estimate the parameters of a linear probability model of D_{fp}^{t+1} (as defined before) as a function of the interaction of similarity S_{fp}^{t} (as defined before) with a metric of skilled labour (number of non-production workers). I also included as regressors the similarity and the number of non-production workers separately, in order to identify in a correct way the differential effect of the number of non-production workers on the relation between similarity and D_{fp}^{t+1}[18]. I estimate the parameters of this model under different specifications that exclude or include the sales and the number of production workers and interactions of these variables with similarity, given their expected[19] and actual[20] correlations with the number of non-production workers. I exclude from all the specifications all those variables that were found to be statistically not significant in the regressions presented in Table 1.

Formally, I estimate the parameters of three different specifications of the following linear probability model:

$$\mathrm{D}_{fp}^{t+1} = \alpha_0 + \alpha_1 S_{fp}^{t} + \alpha_2 nonprodworkers_f^t + \alpha_3 S_{fp}^{t} * nonprodworkers_f^t +$$

[18] I also explored empirically the possibility that each of the three types of capital stock reported in the EAM were correlated with the probability of product addition either by itself or interacted with the similarity. I did not find any statistically significant result in any case

[19] It might be the case that larger firms employ more non-production workers. Analogously, it is reasonable to expect that firms with more non-production workers need more production workers, as long as there exists complementarity between these two groups, as is surely the case.

[20] The correlations between the number of non-production workers and the logarithm of sales and the number of production workers are both above 0.3.

$$\alpha_4 other_f^t + \alpha_5 other_f^t * S_{fp}^t + \gamma_f + \gamma_p + \gamma^t + \epsilon_{fp}^{t+1} \tag{3}$$

S_{fp}^t and D_{fp}^{t+1} are defined as before. γ_f, γ_p and γ^t are fixed effects by firm, product and year, respectively, as in (2). The variable $other_f^t$ can take one of three values (one in each specification): the number of production workers, the logarithm of sales, or none of these two. As for the model in expression (2), I estimate the parameters of the model in expression (3) by using a methodology of least squares with clustered errors by firm.

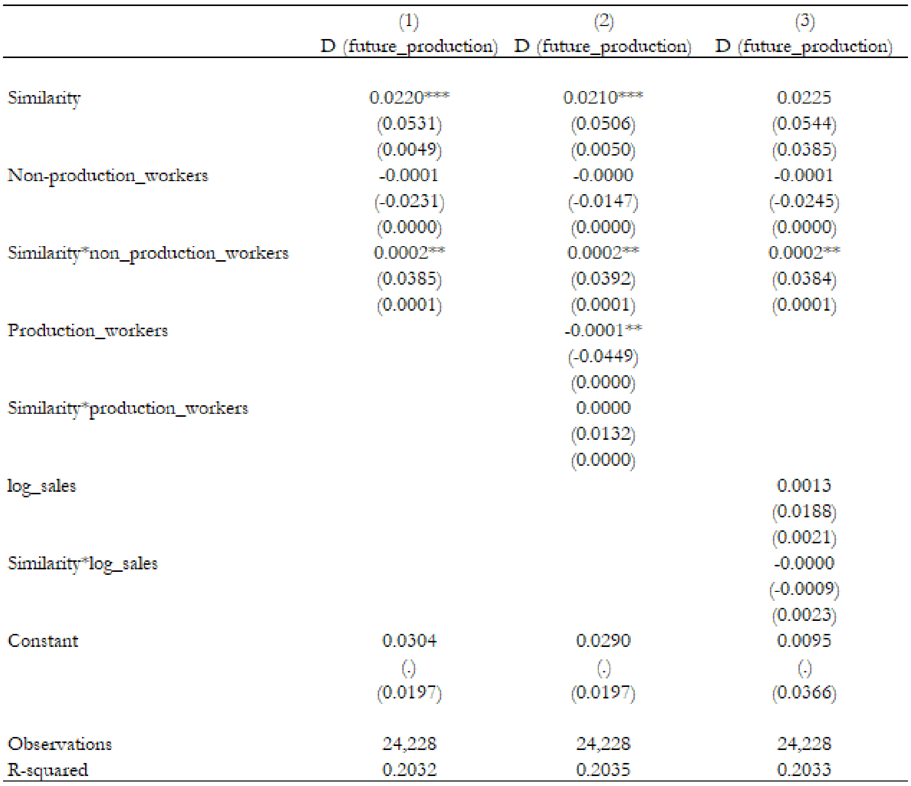
The number of non-production workers is the best available measure for skilled labour in the EAM, although it is admittedly imperfect. As mentioned above, Sun et al. (2020) defines human capital at the firm level (measured here as the amount of skilled labour) as the set of skilled and creative workers who are in charge of innovation, new ideas and knowledge-intensive activities. For this, they take into account several important previous papers that provided similar definitions of human capital. The number of non-production workers in the EAM includes several categories that fit properly into this definition, such as managers, scientists, executives, staff in charge or research and development activities, and the people in charge of conducting the economic policies of firms. However, it also includes other categories such as the people in charge of conducting the financial and administrative policies of firms, and the sales, administrative and distribution staff. Unfortunately, it is not possible to separate these latter from the former, as I explained in detail in the previous section.

The Table 2 presents the results of the estimation of the parameters in expression (3) under different specifications. Its first column shows the results when the only included regressors are the similarity, the number of non-production workers and the interaction of those two variables. The similarity remains significant even at a level of significance of 1 %, and its interaction with the number of non-production workers is also significant at such level of significance. This means that potential new products that require in the present input mixes that are more similar to the input mixes used by the firms have a higher probability of being added one year ahead, and this difference between more and less similar products in terms of their probabilities of addition is larger in firms with more skilled labour. In other words, the relation between the similarity in the input mix and the probability of addition is tighter in firms with more non-production workers.

The columns 2 and 3 of Table 2 include the number of production workers and the logarithm of sales, respectively. In both cases I included both the respective variable and its interaction with similarity, as these interactions are correlated with the interaction of the number of non-production workers with similarity[21]. An alternative and more economic explanation for the inclusion of these interactions is that it is reasonable to hypothesize that larger firms and/or

[21] This because the number of non-production workers is correlated with the number of production workers and the logarithm of sales, as explained in footnotes 14 and 15.

firms with more production workers might be able to maintain their firm-input-specific productivities to a larger extent over time, just as I hypothesized for the skilled labour. Larger firms might have more resources to invest in the activities that are needed to maintain such productivities, and firms with more production workers might be more able to retain skills that might be complementary to the persistence of firm-inputs specific productivities (such as manual proficiency and experience).



	(1)	(2)	(3)
	D (future_production)	D (future_production)	D (future_production)
Similarity	0.0220***	0.0210***	0.0225
	(0.0531)	(0.0506)	(0.0544)
	(0.0049)	(0.0050)	(0.0385)
Non-production_workers	-0.0001	-0.0000	-0.0001
	(-0.0231)	(-0.0147)	(-0.0245)
	(0.0000)	(0.0000)	(0.0000)
Similarity*non_production_workers	0.0002**	0.0002**	0.0002**
	(0.0385)	(0.0392)	(0.0384)
	(0.0001)	(0.0001)	(0.0001)
Production_workers		-0.0001**	
		(-0.0449)	
		(0.0000)	
Similarity*production_workers		0.0000	
		(0.0132)	
		(0.0000)	
log_sales			0.0013
			(0.0188)
			(0.0021)
Similarity*log_sales			-0.0000
			(-0.0009)
			(0.0023)
Constant	0.0304	0.0290	0.0095
	(.)	(.)	(.)
	(0.0197)	(0.0197)	(0.0366)
Observations	24,228	24,228	24,228
R-squared	0.2032	0.2035	0.2033

Table 2. Probability of addition in t+1 as a function of similarity in the use of inputs, its interaction with skilled labour and other variables. Beta coefficients in parentheses one line below the estimators. Standard errors in parentheses two lines below the estimators.

The key fact to be highlighted from the Table 2 is that the interaction of similarity with the number of non-production workers remains significant under all the presented specifications, and its sign and size do not change when other variables that are correlated with it are excluded. In other words, there is strong evidence that potential new products that are more similar to the firms in terms of the input mix in the present are more likely to be added one year ahead, and that this advantage of the more similar products is larger in firms with more skilled labour (non-production workers). This effect of the non-production workers on the derivative of the probability of product addition with respect to similarity is sizeable. Namely, in the model of column (1) an increase of one standard deviation in the number of non-production workers increases such derivative from 5.3 percentage points per change of one standard deviation in similarity to 9.1 percentage points per change of one standard deviation in similarity. This sizeable effect does not change much from the model of column (1) to the models of columns (2) and (3).

Fact 4: The positive correlation between the probability of product addition by the Colombian manufacturing firms and the interaction of the input mix with skilled labour persists over time.

So far I have found that different potential new products have different average probabilities of being added by the Colombian manufacturing firms, that the potential new products with an input mix more similar to the input mixes used by those firms are more likely to be added by them in the near future, and that this correlation between the similarity in the input mix in a year and the probability of product addition one year ahead is stronger in firms with more skilled labour (measured by non-production workers). These facts are consistent with the three hypothesis stated at the beginning of this section, in the sense that if these latter were true, one should observe such facts.

However, it is possible to analyze additional evidence in order to establish more (and stronger) facts in favor of hypotheses 2 and 3. Given some key features of the EAM, it is possible to analyze if the similarity in a specific year is correlated with the probability of product addition not only *one year ahead*, but also several years ahead. If this were true, this finding would constitute stronger evidence in favor of Hypotheses 2 and 3.

If a potential new product with a higher similarity to firm f in t is still more likely to be added k years ahead by f, this might be interpreted at the light of Hypothesis 2 as evidence that the firm-input-specific productivities of f in t remain active even k years ahead to some extent [22], and not only one year ahead, as I conjectured from the Fact 2 of this section. This persistence would explain that the products produced in t were similar to those added in $t + k$ in their input mixes. Under this interpretation, f would choose to produce both in t and $t + k$ those products that require intensively the inputs for which it has high productivities. As productivities in t and $t + k$ would coincide to some extent under this hypothesis (because they persist over time to some extent), the produced products in both periods should be similar in terms of their input mix (which is the feature that motivates their selection, given the input-specific productivities of f).

If I additionally found that the possible correlation between probability of addition in $t + k$ and the similarity in t is stronger in firms with more skilled labour, this would constitute stronger (and better) evidence in favor of Hypothesis 3. If even k years ahead skilled labour remained statistically significant as a booster of this correlation, this might be interpreted as suggestive evidence that firms with more skilled labour in t have an advantage to maintain their

[22] By "active" I mean that if a firm has a high productivity at using a specific input in t, it still has a high productivity at using it k years ahead.

firm-input-specific productivities to a larger extent not only one year ahead (as I conjectured from Fact 2), but also k years ahead.

In order to establish this possible additional evidence in favor of hypotheses 2 and 3, I estimate the parameters of linear probability models that are very similar to the model that corresponds to the first column of Table 2, with the main difference that the dependent variable is not anymore a dummy that indicates if a potential new product in t is produced or not in $t+1$. Instead, I use in each case a dummy D_{fp}^{t+k} (for $k = 2...5$) that equals one if a potential new product in t is produced in $t+k$, and zero otherwise. This dummy is exclusively defined by the production or not production of the respective product in $t+k$. In other words, it equals one if the respective potential new product is produced in $t+k$, no matter if it was produced or not in the preceding years between t and $t+k-1$[23], and zero otherwise.

The other important difference is that I include here as regressor a dummy $Dropped_f^t$ that equals one if the firm f drops at least one product from its product mix in t and zero otherwise. This variable attempts to capture the possible fact that firms require firm-specific scarce factors to add new products in the long term that cannot be purchased in the market but only built within the firm, as hypothesized by Sutton (2012). If this were the case, the firm might need to drop some products in t to add others later. This variable can also capture possible demand shocks across products because of changing preferences from some products to others within the feasible set of a firm. Namely, consumers might switch from some products to others within the feasible set of a firm, and the firm might react by dropping the products now less preferred by the consumers from their production mixes, and adding the now more preferred products to it. This type of switching is not considered in the hypotheses of this section, but it is worth to take into account its possible existence in this empirical analysis for a span longer than one year, given its possible correlation with the similarity.

Formally, I estimate the parameters of the following linear probability model for $k = 2...5$:

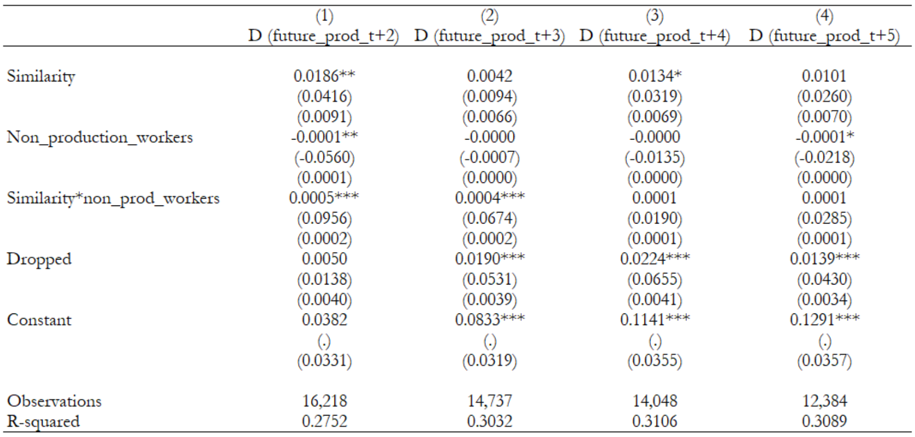
$$\begin{aligned}\mathrm{D}_{fp}^{t+k} = \alpha_0 + \alpha_1 S_{fp}^t + \alpha_2 nonprodworkers_f^t + \alpha_3 S_{fp}^t * nonprodworkers_f^t + \\ \alpha_4 Dropped_f^t + \gamma_f + \gamma_p + \gamma^t + \epsilon_{fp}^{t+k}\end{aligned} \tag{4}$$

The key feature of the EAM that allows me to run regressions to estimate the parameters of the expression (4) for $k = 2...5$ is that I can trace firms across

[23] I ran alternative regressions in which the dummies D_{fp}^{t+k} are defined in a more restrictive way, and my main conclusions did not change. Namely, in this alternative specification the dummies D_{fp}^{t+k} equal one if the corresponding potential new product in t is *never* produced in any year between t and $t+k-1$ and it is produced in $t+k$, and zero otherwise

years with univocal identifying codes, and I can observe every year what products are produced by each firm. In addition, the codes that identify the products remain the same across years for sufficiently long spans (see the previous section for details). All this allows me to construct the set of potential new products for each firm in each year, to identify each year the firms that drop at least one product, and to identify the potential new products of each firm in each year that are added by it in each of subsequent years.

Table 3 shows the results of estimating the parameters of expression (4) for $k = 2...5$ by using least squares with errors $epsilon_{fp}^{t+k}$ clustered by firm. The most important result for this work is that the interaction of similarity with non-production workers remains statistically significant and positive even 3 years ahead. In other words, Colombian manufacturing firms add more often in the future those potential new products that are more similar in t to their production in terms of the input mix and this happens even 3 years later. Moreover, this difference in probabilities in favor of more similar products is larger in firms that employ more non-production workers.



	(1) D (future_prod_t+2)	(2) D (future_prod_t+3)	(3) D (future_prod_t+4)	(4) D (future_prod_t+5)
Similarity	0.0186**	0.0042	0.0134*	0.0101
	(0.0416)	(0.0094)	(0.0319)	(0.0260)
	(0.0091)	(0.0066)	(0.0069)	(0.0070)
Non_production_workers	-0.0001**	-0.0000	-0.0000	-0.0001*
	(-0.0560)	(-0.0007)	(-0.0135)	(-0.0218)
	(0.0001)	(0.0000)	(0.0000)	(0.0000)
Similarity*non_prod_workers	0.0005***	0.0004***	0.0001	0.0001
	(0.0956)	(0.0674)	(0.0190)	(0.0285)
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Dropped	0.0050	0.0190***	0.0224***	0.0139***
	(0.0138)	(0.0531)	(0.0655)	(0.0430)
	(0.0040)	(0.0039)	(0.0041)	(0.0034)
Constant	0.0382	0.0833***	0.1141***	0.1291***
	(.)	(.)	(.)	(.)
	(0.0331)	(0.0319)	(0.0355)	(0.0357)
Observations	16,218	14,737	14,048	12,384
R-squared	0.2752	0.3032	0.3106	0.3089

Table 3. Probability of addition in t+k (for $k = 2...5$) as a function of similarity in the use of inputs and other variables. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

Very interestingly, the dropping of products in t becomes a positive and significant regressor of future addition of products from year 3 on, when the interaction of similarity with skilled labour stops being significant. This combination of results seems to suggest that firms f can add potential new products to their product mixes in the medium term (up to three years later) without dropping products in t as long as such products are sufficiently similar to their production in t and as it has sufficient skilled labour (measured here by non-production workers), but these factors stop being relevant from some point and dropping products in t becomes necessary to add new products in the longer run. The positive sign and high statistical significance of the variable for the dropping of products in t can also be capturing possible correlated demand shocks across products. As explained before, it might be the case that firms drop some products in t because their demands fall, and subsequently add other products in $t+k$ (for $k = 3...5$) because their demands increase. This possibility is not explored to a larger extent in this work, and remains as an important topic for the future.

The number of observations falls monotonically as the span becomes longer because fewer firms survive as I move to longer spans. This might cause a selection bias to some extent if the similarity, its interaction with skilled labour and/or the dropping of products were correlated with the factors that determine the survival of firms. However, I show in the chapter 3 of this thesis that the main conclusions from Table 3 remain valid when only the firms that survive continuously until year $t + 5$ are used.

The results shown in Table 3 are indicative of the possible validity of hypotheses 2 and 3, as explained above. Firms do add more often in a future more distant than just one year ahead those potential new products that are more similar in the present to their current production in terms of the input mix, and this advantage of this type of products is stronger in firms with more skilled labour (measured here by the number of non-production workers). This can be interpreted as suggestive evidence that firm-input-specific productivities in a year t remain to some extent several years ahead, and that this persistence is stronger in firms with more skilled labour in t. In other words, the results shown in Table 3 can be interpreted as suggestive evidence that the economic principles that constitute the hypotheses 2 and 3 (this is, persistence and the effect of skilled labour on such persistence) might actually be true even 3 years ahead.

Fact 5: Some Colombian manufacturing firms perform many more product additions than others.

The main interest of this work (and of the other two chapters of this thesis) is to assess the possibility that the accumulation of firm-input-specific productivities drives the phenomenon of product addition at the micro (firm) level. This is why the bulk of this work is at the level of the firm. However, it is possible to conjecture some possible aggregate facts that one should expect to observe in the data of an economic sector or of an economy as a whole if hypotheses 1 and 2 above were true. One of them is that there should be a high and more than proportional concentration of the effective events of product addition (this is, of cases in which a firm adds a new potential product to its product mix) in some firms. This because hypotheses 1 and 2 jointly imply that firm-input-specific productivities drive the process of product addition, and that such productivities persist over time to some extent. If some firms had higher firm-input-specific productivities than others for several inputs (as is almost surely the case), this conclusion from hypotheses 1 and 2 would necessarily imply that some firms add systematically more of their potential new products to their product mixes than others.

In order to analyze if the product addition events are indeed concentrated more than proportionally in some firms, I calculate the cumulative share of those events between 1992 and 2017 that is performed by each value of the cumulative distribution of the Colombian manufacturing firms. This latter cumulative distribution is organized in such a way that it starts with the firms that performed less product addition events than all the other firms in this period (zero events), and adds subsequently the firms that added more products than this latter but less than the rest of firms, and so on. In one sentence, it goes from the smallest to the largest firms in terms of the number of events of product addition. These two cumulative distributions allow me to draw the ”Lorenz curve-type” graph that is shown in Figure 2. The x axis of this figure measures the cumulative distribution of Colombian manufacturing firms (organized in the way that I explained before) and the y axis represents the cumulative distribution of product addition events.

The main conclusion from Figure 2 is that a share of top firms account for a much more than proportional share of events of product addition. Namely, the 60% of Colombian manufacturing firms that performed less product addition events account for just a bit more than 20% of the total number of events of product addition of the manufacturing sector during the period 1992-2017, whereas the top 20% of Colombian manufacturing firms in terms of product addition performed almost 60% of this total number. This finding is not new in the literature. Freund and Pierola (2016) found a very similar result for a sample of 45 countries. However, it is important for this work because in this context it can be interpreted as suggestive evidence of the possible validity of the hypotheses 1 and 2 at the macro level. This because if firm-input-specific productivities drive the process of product addition and they persist over time to some extent, some firms (those with higher firm-input-specific productivities) should add systematically more of their potential new products to their product mixes, as long as some firms have indeed initial higher firm-input-specific productivities than others (as is almost surely the case).

5 Conclusions

Product addition at the micro (firm) level is a key economic phenomenon. This work presents a set of novel empirical findings of this phenomenon. For this, I exploit a very detailed dataset of Colombian manufacturing firms that includes product-level and input-level modules for every Colombian manufacturing firm with more than 10 employees or sales above 120k U.S. dollars (in 2021), in addition to several other variables such as sales, capital stock and labour. I construct this dataset for the period 1992-2017, which includes several economic cycles. The dataset includes a firm identification code, which allows me to trace firms across time. These features allow me to explore patterns of product addition and its possible correlations with several variables at the firm level.

To start, I find that the conditional probabilities of addition of manufacturing products by Colombian firms are different for different products. Namely, I found that some products have much higher empirical probabilities of addition than others, just as several other authors have found for other countries. Secondly, I establish the empirical fact that Colombian manufacturing firms add more often in the near future those of their potential new products

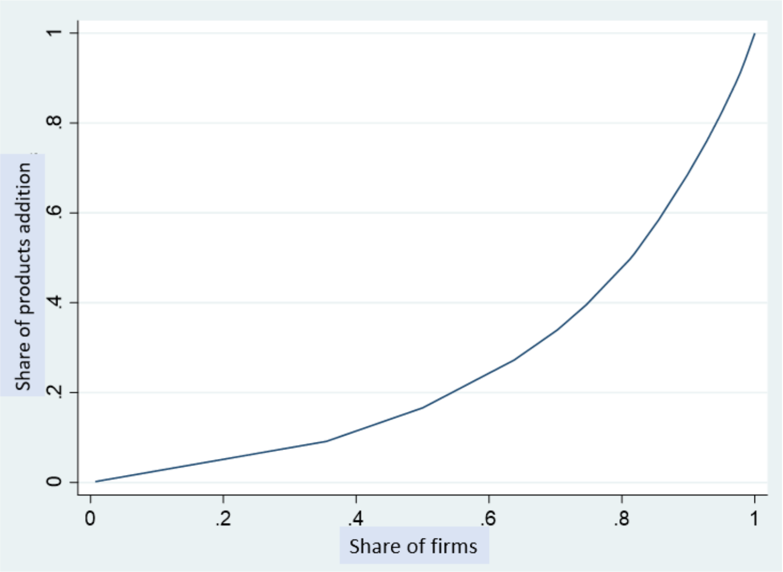


Figure 2. Cumulative distribution of the total number of product additions across Colombian manufacturing firms

that are intensive in the same inputs that they use intensively in the present. Thirdly, I find that this correlation between similarity in the input mix and the probability of product addition is stronger in firms with a higher number of non-production workers (which I use as a measure of skilled labour). In other words, the advantage of more similar products (in terms of the input mix) is more pronounced in firms with more non-production workers. Fourthly, I find that this skilled-labour-augmented relationship between similarity in a specific year and the probability of addition of potential new products in the future remains positive and significant even three years ahead. Finally, I find that the top Colombian manufacturing firms in terms of their capacity to add new products account for a disproportionately high share of product addition events of the Colombian manufacturing sector.

I interpret this set of findings as suggestive and non-conclusive evidence of the possible existence of persistent firm-input-specific productivities that determine the phenomenon of product addition to some extent and that persist over time to a larger extent in firms with more skilled labour. This evidence is just suggestive and indicative (instead of strong or conclusive) because in all the cases I was only able to conclude that *if* the corresponding hypothesis were true, *then* one should observe the corresponding empirical fact. In other words, I was only able to establish *necessary* evidence in all cases, but not *sufficient* evidence in any case, as all the facts established here might be explained by the hypothesized firm-input-specific productivities, but they might also be explained by many other characteristics or decisions of the firms. I have not presented here evidence that these productivities do exist and determine the phenomenon of product addition, but just evidence of the existence of facts that one should observe if such productivities existed and determined indeed such phenomenon to some extent.

Establishing sufficient and conclusive evidence that the firm-input-specific productivities exist, have the hypothesized properties and determine to some extent the phenomenon of product addition requires additional work. This is what I do in the two other chapters of this thesis. In the second chapter I formalize the theory for which I want to find empirical validity. Namely, I present a theoretical model of product addition by firms. The properties of such model are consistent with the empirical facts presented in this paper. Its main feature is that the phenomenon of product addition is governed in this model by persistent firm-input-specific productivities, just as I hypothesized here. Most importantly, in the third (and last) chapter I analyze possible evidence in favor of this theory, including causal evidence and evidence of the existence of its main mechanisms and components.

References

Amiti, Mary and Konings, Jozef (2007). "Trade Liberalization, Intermediate Inputs, and Productivity: Evidence from Indonesia", *American Economic Review*, Vol. 97, No. 5.

Bernard, Andrew and Redding, Stephen (2010). "Multiple-Product Firms and Product Switching", *American Economic Review*, Vol. 100, No. 1.

Bernard, Andrew, Redding, S. and Schott, P. (2011). "Multiproduct Firms and Trade Liberalization", *Quarterly Journal of Economics*, Vol. 126, No. 3, 1271-1318.

Bernard, Andrew, Redding, S. and Schott, P. (2007). "Comparative Advantage and Heterogeneous Firms", *Review of Economic Studies*, 74, 31-66.

Boehm, Johannes, Dhingra, S. and Morrow, J. (2019). "The Comparative Advantage of Firms", *CEPR discussion papers*, 13699.

Broda, Christian and Weinstein, D. E. (2010). "Product creation and destruction: Evidence and price implications", *American Economic Review*, 100(3):691–723.
Cohen, Wesley M. and Klepper, S. (1996), "Firm Size and the Nature of Innovation within Industries: The Case of Process and Product RD", *The Review of Economics and Statistics*, Vol. 78, No. 2, 232-243.

Ding, Xiang. (2020), "Industry Linkages from Joint Production", *Unpublished.*

Eaton, Jonathan, Eslava, M., Kugler, M. and Tybout, J. (2011). "Export Dynamics in Colombia: Firm-Level Evidence", *NBER Working Papers*, 13531.

Feenstra, Robert C. and Ma, H (2007). "Optimal Choice of Product Scope for Multiproduct Firms under Monopolistic Competition", *NBER Working Pa-*

pers, 13703.

Feenstra, Robert C. and Rose, A (2000). "Putting Things In Order: Trade Dynamics And Product Cycles", *The Review of Economics and Statistics*, MIT Press, Vol. 82(3), 369-382.

Freund, Caroline and Pierola, Martha (2016). "The Origind and Dynamics of Export Superstars", *IDB Working Papers Series*, IDB-WP-741.

Guo, Diyue. (2019), "Multiproduct Firms and the Business Cycle", *Working Papers Wang Yanan Institute for Studies in Economics (WISE), Xiamen University*, 2019-05-01.

Hausmann, Ricardo and Hidalgo, C. (2009). "Country Diversification, Product Ubiquity, and Economic Divergence", *CID Working Paper, Harvard University*, RWP10-045.

Huergo, Elena and Jaumandreu, Jordi (2004). "How Does Probability of Innovation Change with Firm Age?", *Small Business Economics*, 22, 193-207.

Kamien, Morton I. and Schwartz, N (1975). "Market Structure and Innovation: A Survey", *Journal of Economic Literature*, Vol. 13, No. 1, 1-37.

Klepper, Steven (1996). "Entry, Exit, Growth, and Innovation over the Product Life Cycle", *American Economic Review*, Vol. 86, No. 3, 562-583.

Klette, Tor Jakob (1994), "RD, Scope Economies, and Plant Performance", *The RAND Journal of Economics*, Vol. 27, No. 3, 502-522.

MacDonald, J. (1985), "RD and the Directions of Diversification", *The Review of Economics and Statistics*, Vol. 67.

Nadiri, M. Ishaq (1993). "Innovations and Technological Spillovers", *NBER Working Papers*, 4423.

Ornaghi, Carmine (2006). "Spillovers in Product and Process Innovation: Evidence from Manufacturing Firms", *International Journal of Industrial Organization*, Vol. 24, No. 2, 349-380.

Sun, Xiuli, Haizheng Li and Vivek Ghosal (2020). "Firm-level human capital and innovation: Evidence from China", *China Economic Review*, Vol. 59.

Sutton, John (2012). "Competing in capabilities. The Globalization Process", *Oxford University Press.*

Vernon, Raymond (1966). "International Investment and International Trade in the Product Cycle", *The Quarterly Journal of Economics*, Vol. 80, No. 2, 190-207.

Product addition and comparative advantages in Colombian manufacturing firms in a context of trade liberalization

Andres Trejos

June 19, 2025

Abstract

This work presents new empirical evidence of possible firm-level and firm-product-level determinants of the important phenomenon of product addition at the firm level. I do this by using a stratified sample of a rich panel of Colombian manufacturing firms from 1992 to 2017. Importantly, this panel includes all the firms with 10 or more employees and/or sales above 120k USD per year. I find that firms tend to add to their product mixes products with similar input requirements (input mixes hereinafter) as products already in their portfolios. In addition, I find in general that this correlation between product addition and similarity in the input mix is stronger on average in firms with more skilled labour (proxied here by the number of non-production workers), especially after controlling for a possible omitted variable bias. These findings are consistent with the theoretical hypotheses presented in the second chapter of this thesis. In addition, I find that the correlation between future product addition and the current similarity in the input mix remains positive and significant even 5 years ahead, but I do not find evidence that less similar products are added each year as I move further into the future, as I hypothesized in the second chapter of this thesis. In addition, I use product-specific exogenous unilateral reductions of tariffs to imports from the U.S. implemented in 2011 by the Colombian government to analyze possible causal evidence in favor of the model presented in the second chapter of this thesis, which I cannot establish unambiguously. Finally, I explore possible empirical evidence in favor of the most relevant mechanism of this model and of its main assumption. Its main mechanism is that those products that are intensive in the inputs in whose use a firm is proficient are more profitable for this firm. Its main assumption is that the firm-input-specific productivities of a firm (which measure its proficiency to use the different inputs) grow over time as this firm uses the respective inputs to a larger extent. The empirical evidence is mostly consistent this mechanism and assumption.

1 Introduction

Product addition at the firm level is a key phenomenon for economies to achieve economic efficiency and growth. As I explained in detail in the first chapter of this thesis, this phenomenon has been found to be the main driver of the process of product diversification at the firm level, which in turn is an important source of reallocation of resources within firms towards their more efficient use, as it allows firms to move their resources to more profitable activities (by switching products). Given this, product diversification contributes as much as the phenomenon of firm entry and exit to the evolution of U.S. aggregate manufacturing output. It is also an important determinant of variability in the aggregate economic activity [1]. In spite of this, the determinants of the process of product addition at the firm level have not been sufficiently analyzed. This paper contributes to closing this gap by presenting new facts about the possible determinants of this process for the case of Colombian manufacturing firms.

By using a very rich dataset of Colombian manufacturing firms, this paper does three things. Firstly, it explores the possible nonlinear relationship between product addition at the firm level and several firm-level and firm-product-level possible determinants. Namely, I explore the possible empirical validity of the first theoretical proposition from the second chapter of this thesis that those potential new products that require an input mix that is more similar to that used in a year by a firm are more easily added by this latter in subsequent years. Moreover, I also explore the possible empirical validity of the second theoretical proposition in the second chapter of this thesis that this relation between product addition and similarity is larger in firms with more skilled labour, and that firms with more skilled labour add more easily potential new products to their product mixes (all else constant).

Secondly, I explore the possible correlation between product-level costs of production and the probability of product addition. For this, I use as instruments the tariffs imposed by the Colombian government to the imports from the U.S. during a generalized and unilateral reduction of these tariffs. As this generalized reduction in tariffs is exogenous to the decisions of product addition of the Colombian manufacturing firms and it is precisely costs of production what cause the role of similarity as a determinant of product addition in the theoretical model of the second chapter of this thesis [2], this exploration constitutes an examination of possible causal evidence in favor of the theoretical model presented in the second chapter of this thesis. For this analysis I use the same dataset of Colombian manufacturing firms mentioned before.

Thirdly, I explore the possible empirical validity of the main mechanism of the theoretical model from the second chapter of this thesis, and also of its main assumption. Namely, I explore possible evidence that the interactions between input-output coefficients and firm-input-specific productivities determine the process of product addition by firms. In order to analyze the empirical validity of this mechanism, I explore the possible correlation between the probability of addition and the aforementioned interaction. For this analysis I use the same dataset of Colombian manufacturing firms mentioned before. As for the main assumption (which I named "learning by using in the second chapter of this thesis), I explore the possible correlation between the use of inputs by firms in

[1] Please see the first chapter of this thesis for a detailed explanation of the importance of product diversification and for the references on which this explanation is based.

[2] Please see the third theoretical proposition in the second chapter of this thesis for details.

a given year and the corresponding firm-input-specific productivities one year ahead.

As for the first task mentioned above, I find a positive and statistically significant correlation between similarity in the use of inputs and product addition one year ahead. I also find that such correlation is stronger in the firms with the highest quantities of non-production workers (which I use as a proxy of skilled labour). Very importantly, my results for the effect of non-production workers on the correlation between similarity and product addition improve when I include the potential cost of production as a regressor. This suggests a possible omitted variable bias in the specifications that exclude this variable. In addition, I find that the positive correlation between product addition and similarity remains significant for longer time spans, although its size does not decrease monotonically as I move further away in the future, as expected. In summary, this paper presents evidence that is mostly consistent with the first proposition in the second chapter of this thesis, and partially in favor of its second proposition (especially when the potential cost of production is included as a regressor).

As for the second task mentioned above (called here "causality analysis"), the results presented here do not conclusively establish or deny causal evidence of the validity of the theoretical model presented in the second chapter of this thesis. This because I explore several specifications with and without instrumented regressors, and the conclusions are heterogeneous (although I do find in all cases a negative point estimator for the cost of production as a regressor for the probability of addition as dependent variable, as expected). Finally, I find empirical evidence that is mostly consistent with the possible existence of the main mechanism of the model presented in the second chapter of this thesis, and also evidence that is mostly consistent with the validity of its main assumption.

It is worth to mention that I perform in this paper a structural estimation of the structural input-output coefficients and of the firm-input-specific productivities, which are then used to carry out the second and third tasks mentioned above. Namely, I use them for the causality analysis and for the empirical examination of the main mechanism and the main assumption of the theoretical model presented in the second chapter of this thesis. For this, I use the first-order conditions of the product-specific conditional profits of a firm with respect to the different inputs, which I derived in the second chapter of this thesis. A log-linearization of these conditions yields a linear estimable expression in which both the input-output coefficients and the firm-input-specific productivities are identified.

This paper is closely related to the very recent and far from conclusive branch of literature that analyzes empirically the possible determinants of the phenomenon of product addition at the firm level. The papers that belong to this branch share two key features with the work presented in this paper. Firstly, their authors test the empirical validity of specific theoretical predictions from theoretical models that are proposed by themselves, instead of performing agnostic regressions in the search for possible significant correlations. Secondly, their authors carry out a causality analysis or a complementary empirical analysis that attempts to establish empirical evidence in favor of their main mechanisms. Very importantly, the conclusions from these works are contradictory to some extent, as I explain below. This highlights the importance of the work presented here, as it brings new evidence into this nascent discussion and sheds light about possible mechanisms that might lie behind some results from the past.

Ding (2020) used U.S. data to test his prediction that it is similarity in the use of knowledge inputs (and not in the general use of inputs) what matters for product addition, and to test for the validity of the main mechanisms involved in his model. In contrast, Boehm et al. (2019) used Indian data to test for their prediction that the probability of product addition is higher for more similar products in terms of the *general* use of inputs (and not just of the use of *knowledge* inputs), and to test for the empirical validity of the main mechanisms involved in their model. They also used the exogenous process of dereservation in India to establish causal evidence in favor of their model.

My work is very similar to the one carried out by Boehm et al. (2019) both in terms of the nature of the datasets and of the direction of the results, as they also explored the possible existence of an empirical positive correlation between similarity in the input mix and product addition, and they also found evidence of this existence. However, the analyses presented here explore the role of additional characteristics of firms in the process of product addition. Namely, I explore both the empirical correlation between product addition and skilled labour, and also the way in which skilled labour determines the size of the correlation between similarity in the input mix and product addition. In short, I find that the amount of skilled labour used by a firm determines the size of the correlation found by Boehm et al. (2019) and by myself between product addition and similarity in the input mix. This role of skilled labour might be actually related to Ding's results, as it might be indicative of the importance of knowledge inputs (which are closely related to skilled labour). I do not explore this possibility in this paper.

Given the aforementioned importance of the process of product addition at the firm level for growth and efficiency, identifying its determinants is important for growth policies. This because growth policies should ideally aim at improving the state of firms in the dimensions that are identified as significant determinants of product addition, as this latter phenomenon is in turn a source of growth and efficiency. As this work identifies some of these determinants, its conclusions may be used to inform the process of design and implementation of growth policies. This is why I include a brief discussion of the main possible policy implications of the findings of this paper in the last section.

This document has seven sections. The first of them is this introduction. The second section presents a brief description of the theoretical model proposed in the second chapter of this thesis and states its three propositions, as I refer to this model and these propositions often throughout the rest of this document. The third section describes in detail the dataset to be used in the subsequent sections. The three subsequent sections present the results of the first, second and third tasks described above, respectively. Namely, the fourth section presents all the results of the regressions of product addition on similarity in the input mix, skilled labour and the interactions between these two latter. The fifth section presents the results of the causality analysis. The sixth section presents the results of the analysis of the empirical validity of the main mechanism and the main assumption of the theoretical model presented in the second chapter of this thesis. Finally, the seventh section presents the conclusions of this work and its main possible policy implications.

2 The theoretical model and its propositions

I propose in the second chapter of this thesis a static model of partial equilibrium with dynamic implications. In this model all firms are potentially multiproduct. Each firm maximizes its profits in a year t by choosing its product mix, its produced quantity of each produced product, the quantities of labor and capital to be used in its production process and the quantities of all the inputs to be used for the production of each produced product [3]. Very importantly, this use of inputs in t determines the firm-input-specific productivities in $t+1$. More specifically, there is a process of *learning by using* in which a firm f's firm-input-specific productivity to use a particular input k grows more between t and $t+1$ the more k is used by f in t. This implies that there is persistence in the firm-input-specific productivities in this model. These latter do not only not disappear from one year to other, but they may grow as a consequence of the inputs use.

Given this structure, this theoretical model yields naturally three key results. Firstly, a firm f adds more easily in the future those potential new products that require input mixes that are more similar to the input mix that f uses in the present. This result comes from the fact that f finds it more profitable to produce in t those products that require intensively those inputs for whose use f has a high productivity (let us call these group "f's preferred inputs"). In turn, the production of these effectively produced products requires intensively the use of f's preferred inputs, as they are by definition intensive in these latter. This intensive use of f's preferred inputs makes f even more productive at using f's preferred inputs in $t+1$ (by the process of learning by using), which implies that f will choose once again in $t+1$ potential new products that are intensive in f's preferred inputs to be added to its product mix (given their higher profitability for f).

As a consequence of the dynamics described above, there is a higher similarity in the intensity in different inputs (input mix) between the products produced in t and the new products produced in $t+1$ than between the former and the potential new products *not* produced in $t+1$. This because the new products produced in $t+1$ are intensive in f's preferred inputs (just as the products produced by f in t), whereas the potential new products *not* produced

[3] Here the term "inputs" does not make reference to capital or labour, but only to the other physical inputs used in the production process that are directly transformed into outputs. Please see the first chapter of this thesis for a detailed explanation of this differentiation and of its relevance.

in $t+1$ are not intensive in f's preferred inputs (that is precisely why they are not chosen to be produced). As f's productivities to use its preferred inputs keep on growing in time, this higher similarity for the produced potential new products than for the not produced ones persists in all the subsequent years. In short, firm-input-specific productivities determine the profitability to produce different products. As these productivities remain in time, the products produced in different years are similar to each other in terms of their input mixes (as they are all intensive in the same inputs).

This relationship between product addition and similarity changes over time in the model presented in the second chapter of this thesis. If a product p's input mix is not sufficiently similar in t to the input mix used by a firm f in t as to be added to f's product mix in $t+1$, it may be eventually added in a subsequent year, all else constant. This because firm-input-specific productivities may grow mechanically over time in this model (although they grow to different extents, depending on how intensively they are used). In other words, all the products may eventually be added (sooner or later) by all firms to their product mixes, all else constant. However, how soon a product p is added to a firm f's product mix depends positively on how similar is p's input mix to the input mix used by f in t. Namely, the less similar products are added later, as I explain below.

The less similar products may be added later by a firm f to its product mix because it may take longer to f to reach the firm-input-specific productivities that are needed to produce these products. This happens because of two reasons. Firstly, f is not proficient in t in the inputs that are intensively used to produce p (if it were, it would use these more intensively and its input mix would be more similar to p's one in t). This implies that f does not use those inputs intensively in t. This low intensity in the use of these inputs in t makes the increase from t to $t+1$ in f's productivity to use them less pronounced than the increase expected for other inputs. Secondly, this latter fact makes the firm once again less intensive in these inputs than in others in $t+1$, and so on. As a consequence, less and less similar products are added to the product mix of a firm every year.

All the pieces of intuition presented above are summarized in the first proposition of the second chapter of this thesis, which states the following[4]:

Proposition 1 (main proposition): *Firms are more likely to add in the future potential new products whose input mixes are closer to the firms' current input mixes. The further one moves away from the present time into the future, the less similar are the input mixes of the potential new products that are added to firms' product mixes.*

As I explained above, I assume that the process of learning by using is more pronounced in firms that use more skilled labour. Formally, the increase between t and $t+1$ in f's productivity to use an input k that occurs when k is

[4] Please see the second chapter of this thesis for a formal proof of this proposition and of the propositions 2 and 3 below, and for a formal definition of potential new products and of similarity in the input mix

more used in t is higher if f uses more skilled labour in t[5]. As a consequence of this, two results arise. Firstly, all the potential new products are more easily added in all the years after t by firms with more skilled labour. Intuitively, these firms exhibit larger increases in all the firm-input-specific productivities between all the pairs of contiguous years from t on. Therefore, all the products are more profitable for them in all the years.

Secondly (and more importantly), this advantage of the firms with more skilled labour favors to a larger extent the products that require input mixes that are more similar to the input mix used by those firms. In other words, the advantage of the more similar products stated in the proposition 1 above is more pronounced in firms with more skilled labour.

The important fact stated in the previous paragraph comes from the key feature of the model proposed in the second chapter of this thesis that skilled labour interacts *multiplicatively* with the use of each input in t in the equation that characterizes the process of learning by using, which determines the firm-input-specific productivities in $t+1$. This multiplicative nature of this interaction implies that the skilled labour boosts the productivity of a firm f to use a particular input k in the future to a larger extent if f is used to a larger extent by f in t. As f's preferred inputs are used to a larger extent by f in t, these inputs are more favored (in terms of f's productivities to use them) by the skilled labour's effect on the process of learning by using. This differential effect in favor of these inputs increases to a larger extent the future profitability of the potential new products that require them intensively. As I explained above, these latter products are the ones effectively produced by the firm in the future, and they have a higher similarity with f in terms of the input mix.

These two results from the model presented in the second chapter of this thesis are summarized and materialized in the following proposition:

Proposition 2 (the role of skilled labour): *Firms with more skilled labour are more likely to add new products in the future. The advantage of the more similar products stated in Proposition 1 is greater in firms with more skilled labour.*

The theoretical model presented in the second chapter of this thesis also predicts two things about the effects of a generalized reduction in the prices of inputs in the present (year t). Firstly, there occurs a heterogeneous increase in the profitability of addition of the different products by the different firms. More specifically, this increase is higher for the firm-product combinations for which the corresponding products intensively require for their production those inputs whose prices fall more. Secondly, this compound differential effect is even larger if the corresponding firms have high firm-input-specific productivities in the inputs whose prices fall more.

The intuition for the second prediction above is as follows: assume that a

[5] Please see the second chapter of this thesis for an explanation of the soundness of this assumption.

firm f is better at using (say) glass than (say) paper and the opposite is true for another firm $f2$. Also assume that the price of glass falls more than that of paper in t. Given its proficiency to use glass, f is able to take more advantage of the pronounced reduction in the cost of glass in t than $f2$, and f uses glass to a larger extent in this year. As a consequence, f's productivity to use glass in subsequent years increases more than that of $f2$ from t to $t+1$ by the process of learning by using. In turn, this increases f's profitability to produce glass-intensive products in all the subsequent years to a larger extent than for other firms [6].

As for the intuition for the first prediction, it is as follows: if a product p is intensive in the inputs whose prices fall more, its cost of production falls more than that of the other products in t, and therefore the profitability of its production in t increases more than that of other products. Given this, more firms produce p in t than other products, all else constant. As a consequence, the inputs that are used intensively to produce p (this is, the same whose prices fell more) are demanded to a larger extent in t than other inputs, all else constant. This increases the productivities of firms to use those inputs in the future more than the productivities to use other inputs, because of the process of learning by using. In turn, this increases the profitability of producing p in all the subsequent years to a larger extent than that of producing other products, all else constant.

I summarized all these compound effects in the second chapter of this thesis in the following proposition:

Proposition 3 (*Effect of exogenous changes in the prices of inputs*): *A decline in the cost of an input k leads firms to add potential new products that use this input intensively. This effect is greater for firms with a higher firm-input-specific productivity to use k.*

3 Description of the database

This work uses a very detailed and comprehensive database for the Colombian manufacturing sector. It is often called EAM, which stands for its initials in Spanish (*"Encuesta Anual Manufacturera"*). I will use these initials throughout this document. The Administrative Department for Statistics of Colombia -DANE- started gathering information about the manufacturing firms of Colombia in 1950. Nowadays, it surveys each year every Colombian manufacturing firm with more than 10 employees and/or sales above an amount in Colombian pesos equivalent to 120k U.S. dollars in 2021, and also a representative sample of the smaller manufacturing firms.

The EAM contains several very important modules of information that played a key role in the process of constructing all the variables that were used

[6] Mathematically, this complementarity comes from the fact that the prices of inputs interact multiplicatively with the firm-input-specific productivities and the input-output coefficients. Please see the second chapter of this thesis for a detailed explanation of this.

in the regressions that I present in the subsequent sections of this paper. One of them is a module of products. It includes detailed information of every product sold by each manufacturing firm in each year. This information includes the unitary price of each product produced by each firm in each year, and also the quantity sold of each of them. It also includes both of these variables (unitary price and sold quantity) for exports. In addition, the EAM contains a module of inputs[7], which contains the purchased quantity and the unitary price of every input used by each firm in each year. This module also contains imported quantity and unitary price of imports for every imported input by each firm in each year.

The EAM contains also several firm-level modules. One of them includes detailed information of the employed workers. This module includes the separate number of workers by the type of work they perform. In some years this disaggregation was more detailed than in others. Namely, for some years it is possible to know separately the number of people in sales force, the number of managers, the number of administrative staff, the number of professional staff working in plant production and the number of plant operators. Unfortunately, the information of workers is less disaggregated for other years. As a consequence, I had to include all the workers in only two categories in every year, with the purpose of having consistency across years: production workers (which includes plant operators and professional staff working in plant production) and non-production workers (which includes managers, administrative and sales staff, and very importantly for this work, research and development staff, if these activities are performed by the firm). The EAM includes the total wages paid to workers in each of these categories.

The EAM includes some cost modules that contain relevant information that is not being used for this paper, but many researchers could be interested in using in the future. Namely, these costs modules include information of costs of energy and water, telecommunication services, taxes, interests and rent, among many other variables. In some years it includes complete modules about energy use, which intend to characterize the evolution of main energy sources of firms.

Very importantly for this research, the EAM includes several variables that allowed me to identify each firm and its production in many senses. To start, each firm and plant is uniquely identified with a numerical code that remains the same every year, as long as the firm is surveyed (which depends on the sales and number of employees, as explained before). This allowed me to trace each firm in time.

I used information of the EAM from 1992 to 2017, as information of years before 1992 has several problems and its organization and cleaning would have delayed this research substantially. As for the final year, 2017 was the last available year in the database when the bulk of this research was carried out. All the products and inputs are uniquely identified by a code of the International

[7] Please see the second chapter of this thesis for a detailed definition of "inputs". In short, this term makes reference to all the physical materials used in the productive process that are *directly* transformed into products. As capital and labour *are not* directly transformed into products, these are not included in this category of "inputs" in this thesis.

Standard Industrial Classification (ISIC) revision 2 adapted for Colombia from 1992 to 2000, by a code of the Central Product Classification (CPC) version 1.0 from 2001 to 2012, and by a code of the CPC version 2.0 from 2013 to 2017. In all cases, the EAM uses the most disaggregated levels available of each classification (8 digits for the ISIC rev. 2 and 9 digits for the CPC).

The fact that the EAM uses always the most disaggragated available codes to identify products and materials is very convenient for this research, as it ensures that a code truly identifies a specific product or material to the largest possible extent. However, this also prevented me from performing analysis that required tracing products or inputs across years with different classifications, as there do not exist concordance tables for these levels of disaggregation, but for much more aggregated categories of products. I could have used such levels of aggregation to gain traceability, but I would have lost specificity in the definition of products, which I valued more. As a consequence, I can only trace products within the periods 1992-2000, 2001-2012 and 2013-2017, which fortunately are not extremely short.

The modules of products and inputs are crucial for this work. Unfortunately, they are not publicly, as their free use could violate some Colombian laws that protect the privacy of information for some firms that might be identified even though firms and plants are anonymised. A good example is Reficar, the largest oil refinery of Colombia [8]. Reficar is located in the city of Cartagena. It produces a high share of the total amount of gasoline and diesel used in the country. Even though it is impossible to identify Reficar by its legal ID (as this latter is different from the identifier that was assigned to it in an anonymization process carried out by DANE), it is possible just to search for manufacturing plants in Cartagena in the business of refining, and there will be just one plant with sales as high as to correspond to Reficar. This would allow anyone to see sensible information such as unitary prices by product.

Because of the reasons explained in the last paragraph, the modules of products and inputs can only be accessed from an office located at DANE's main building. Unfortunately, it is not possible to guarantee that regressions that are left running by external researchers run in nighttime and for several days, as interruptions use to happen in the main system. For this reason, I had to use a random sample of firms, as using all firms with all their modules of products and inputs in all years would have implied processing times that exceed by far the time I was allowed to stay in their office. The pandemic of Covid-19 exacerbated this problem, as DANE's main building was closed for almost a year, and then reopened gradually at a very slow pace, with times as restrictive as only 8 hours per week for several months.

The total number of firms included in the EAM has been approximately 8.000 per year in the last two decades. As I am using data from 1992 to 2017, this means that my full firm-year-level database has in total approximately 208.000 observations (8.000 firms per year multiplied by 26 years). As each firm produces on average 3.5 products in a year and uses on average 25 mate-

[8] It has the capacity to refine 150.000 barrels of crude oil per day.

rials in a year, my firm-year-product-level database has in total approximately 728.000 observations, and my firm-year-material-level database has in total approximately 5.2 million observations.

The numbers shown at the end of the previous paragraph reveal the magnitude of the problem. As I will explain in the next section, I had to calculate for this work similarities between each firm and all its potential new products in each year [9] in terms of the use of inputs (or input mix), as such similarity is used as explanatory variable in all the regressions presented in the subsequent sections. This required the calculation of many dot products of two vectors of use of inputs for each product-firm combination for each year. This means that several procedures would have to have been applied to the 5.2 million observations if I had used my full database. I tried this, and unfortunately it exceeded the capacity of the computers available at DANE's office.

In order to bypass this problem, I took a stratified random sample of 2 percent of all the available firms in my full dataset, this is, 160 firms. Strata are defined by an interaction of two dimensions: quartiles of sales of firms in their initial year and number of produced products by firms in their initial year. This procedure guarantees that the sample includes firms that had different sizes in sales and different product scopes when they were born (or started complying with the sales and/or employment requirements to be included in the EAM). I selected these two variables because it is reasonable to expect that they have more influence on the process of product addition than the rest of firm-level variables for which I have information. More specifically, it is reasonable to expect that larger firms add more products to their product mixes, and also that firms that produced more products in their initial year add more products to their product mixes. Even though I do not analyze in depth the relationship of these variables with product addition, the possibility that this correlation exists explains their use as stratification variables.

The fact that each firm's probability of inclusion depends on the two aforementioned variables at their initial year and not on their averages across years prevents me from selecting a sample in which firms that started being small and grew both in sales and in product scope were over-represented. The number of firms selected from each strata is proportional to the share that each strata represents of the total number of firms. This ensures that all the strata are represented in the sample to an extent that is proportional to their importance in the total database of firms.

The initial sales and the initial number of products are correlated [10], which is not surprising, given that they were both chosen as stratification variables because of their possible relationship with product addition. This correlation reinforces the convenience of using them both as stratification variables, as this "bivariate" stratification guarantees that the firms *within* each quartile of sales

[9] Please see the second chapter of this thesis for a formal definition of potential new products. In short, this firm-year-specific set includes all the products that are feasible for the corresponding firm in the corresponding year, but are not produced by it in such year

[10] The correlation between the number of products in the initial year and the sales in that year is nearly 0.55.

are selected in such a way that the firms with different numbers of produts are represented proportionally to their importance in the respective quartile of sales.

This sampling procedure yields a firm-year-level dataset with approximately 7.000 observations, a firm-year-product-level dataset with approximately 14.000 observations and a firm-year-material-level dataset with approximately 100.000 observations.

In addition to information of Colombian manufacturing firms, this paper uses information of the tariffs imposed by Colombia to the imports of all products during the period 1992-2017. Namely, I use for each product in each year the average across all origin countries of the tariffs imposed to imports. This data was obtained from the World Bank's World Integrated Trade Solution (WITS). WITS includes several product classifications. I decided to use here information of products as defined by the Harmonized System (HS), as it is for this classification that I found the best possible table of concordances with the classifications of products provided by the Colombian office of statistics (International Standard Industrial Classification (ISIC) revision 2 adapted for Colombia, Central Product Classification (CPC) version 1.0 CPC version 2.0). As I will explain in detail in section 5, this data on tariffs is used to construct instrumental variables that are used in turn to analyze the possible existence of causal evidence in favor of the model proposed in the second chapter of this thesis.

4 Product addition, similarity in the input mix and skilled labour

This section explores possible empirical evidence in favor of the propositions 1 and 2 of section 2. I do this by estimating the parameters of econometric empirical models. I start by presenting the formal expression of the econometric model to be estimated. I then explain in detail how I calculated the dependent variable and the main regressor of this model. After this, I present and analyze key summary statistics of all these variables. Subsequently, I present the results of estimating the econometric model mentioned above under different specifications and for different spans (from $s = 1$ to $s = 5$). Finally, I present the results under a different way of defining the set of firms to be used in the regressions as an important robustness check.

Econometric model and expected results

Propositions 1 and 2 of section 2 predict that a specific product is more likely to be added by a firm in the future if it is more similar in terms of the input mix in the present time, and that this relationship between addition and similarity is stronger if such firm uses more skilled labour. However, a thorough analysis of the chapter 2 of this thesis reveals that the predicted relationship between product addition and similarity does not necessarily grow linearly as the amount of skilled labour increases [11].

[11] A detailed analysis of the second chapter of this thesis reveals that the extent to which

Given this, I use as regressors interactions of similarity with four dummies that are equal to one for firm-year observations that belong to quartiles one, two, three and four of the overall distribution of skilled labour (proxied here by the number of non-production workers[12]), respectively. In this way, I do not assume any functional form for the extent to which more skilled labour increases the size of the relationship between product addition and similarity, while still being able to analyze if this expected increase as a consequence of more skilled labour does occur. The parameter for each of these interactions measures how greater is the average relationship between product addition and similarity for the firms within the respective quartile of non-production workers than this relationship for the quartile whose interaction is excluded (in order to avoid perfect collinearity). As for the quartile whose interaction is excluded (which is always the lowest one), the relationship between product addition and similarity for the firms within it is given by the parameter for the similarity itself (which is also included in the model).

Quartiles of non-production workers themselves are also included as regressors because the proposition 2 of section 2 also predicts that the firms with more skilled labour add new products more easily, all else constant. In addition, this inclusion allows me to prevent the parameters for the interactions with similarity from capturing the separate effect of skilled labour (proxied by the number of non-production workers) on product addition, instead of its effect on the size of the relationship between product addition and similarity.

Formally, I estimate the parameters of the following linear model in order to analyze the empirical validity of the propositions 1 and 2 of section 2:

$$D_{fp,t+s} = \alpha + \sum_{q=2}^{4} \Gamma_q^s D_{qft} + \delta^s S_{fpt} + \sum_{q=2}^{4} \beta_q^s S_{fpt} * D_{qft} + \psi_t + \phi_f + \tau_p + \epsilon_{fp,t+s} \quad (1)$$

$D_{fp,t+s}$ represents here a dummy variable that equals one if a feasible product p that is not produced by the firm f in t is produced by f in $t+s$, and zero if it is not produced in $t+s$. S_{fpt} represents the similarity in t of the input mix needed to produce p with the input mix used by f. On the other hand, D_{qft} represents a dummy that equals one if the value of non-production workers employed by f in t (which is used here as a proxy for skilled labour) belongs to the quartile q of the total distribution of all observed values of non-production workers across all firms and years. ψ_t, ϕ_f and τ_p represent year, firm and product fixed effects, respectively. The four parameters β_q^s, the four parameters Γ_q^s and δ^s have a superscript s because they are specific for each time span. The summations across quartiles are from the second to the fourth quartile because the first quartile is excluded in both cases in order to prevent perfect collinearity.

the relationship between product addition and similarity grows as a firm uses more skilled labour is not characterized by a explicit function

[12] Please see the first chapter of this thesis for a complete discussion about the soundness of using this variable as a proxy of skilled labour

Given the proposition 1 stated in section 2, I expect δ^s to be positive for every $s \geq 1$ and $\delta^s + \beta_q^s$ to be also positive for every $q = 2...4$ and for every $s \geq 1$. In words, I expect that firms in all the quartiles of skilled labour are always more likely to add more similar products. Formally, this means that I expect the derivative of the expected value of $D_{fp,t+s}$ with respect to the similarity to be positive for every quartile of skilled labour.

Given that the proposition 2 in section 2 states that the advantage of the more similar products is greater in firms with more skilled labour, I expect $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$. In words, I expect that increases in similarity are associated with higher increases in the probability of addition as I move to higher quartiles of skilled labour. As the proposition 1 in section 2 also states that less similar products (in terms of the input mix) are added as s increases, I expect the sum $\delta^s + \beta_q^s$ to fall as s increases for every $q = 2...4$, and I also expect δ^s to fall as s increases.

The proposition 2 in section 2 also states that firms with more skilled labour are on average more likely to add new products. Given this, I expect the sum $\Gamma_q + (\delta^s + \beta_q^s)E[S_{fpt}]$ to increase as q increases, for $q = 2...4$, and I also expect $\Gamma_2 + (\delta^s + \beta_2^s)E[S_{fpt}] > \delta^s E[S_{fpt}]$. In words, the latter condition guarantees that the expected value of $D_{fp,t+s}$ for the second quartile of skilled labour is greater than this expected value for the first quartile of skilled labour, for given a span s. The former condition implies that the same is true when the third quartile is compared to the second and first quartiles, and when the fourth quartile is compared to the third, second and first quartiles. As I show below in Table 1, the empirical estimator for $E[S_{fpt}]$ (this is, the sample mean of S_{fpt}) is 0.48. Therefore, I expect $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ to increase as q increases and $\Gamma_2 + 0.48(\delta^s + \beta_2^s) > 0.48\delta^s$, for $q = 2...4$ and for every $s \geq 1$.

Finally, the error term $\epsilon_{fp,t+s}$ is assumed to be normally distributed with mean zero for every $s \geq 1$. These errors are clustered by firm. In other words, they are assumed to be independent across firms, but are allowed to be correlated with each other across time and products within each firm, and none theoretical structure is imposed for such correlation. Instead, such structure is directly estimated from the data.

Definition and construction of variables

The set of feasible products for a firm f (which is needed to establish the domain of the variable $D_{fp,t+s}$ in each case) is defined in a completely empirical way. Namely, it includes all the products that were ever produced by any firm simultaneously with any of the products produced by f in any year.

The variable S_{fpt} represents the similarity in terms of the use of materials between the firm f and the product p in t. It is calculated as a dot product of 2 vectors of expenditure shares on all the available inputs. One of them (x_{ft}) contains the observed shares for a firm f in t. The other (x_{pt}) contains the average of shares across all the firms in the sample that produce a product p in t. This dot product yields a number between zero and one, as it is normalized by dividing its value by the product of the total variability of both vectors. A value of one would indicate that the input mix needed to produce p in t is identical to the input mix used by f in t. In contrast, a value of zero would indicate that they are completely different. Formally, I calculate S_{fpt} by using the following formula, as in Boehm et al. (2019):

$$S_{fp}^{t} = \frac{\sum_{k=1}^{K} x_{fkt} x_{pkt}}{\left[\left(\sum_{k=1}^{K} x_{fkt}^{2} \right) \left(\sum_{k=1}^{K} x_{pkt}^{2} \right) \right]^{1/2}} \tag{2}$$

Summary statistics

The table 1 below shows different summary statistics of the variables that are used in this section to estimate the model described in (1) for different time spans. Some variables are observed at the firm-product-year level, whereas others are observed just at the firm-year level. In all the cases I used the sample described in detail in section 3.

The variables named *D(future production)* are dummy variables that equal one if a feasible product is produced and zero otherwise [13]. They can indicate in a year t the production of a given product in the current year or k years ahead, for $k = 2...5$.

For each time span I present the summary statistics for *D(future production)* in two different cases. In the first case (*"All products"*) the summary statistics are calculated by using the *future production* dummies of *all* the products in the whole feasible set of the corresponding firm. Formally, the means estimate in this case the *unconditional* probabilities of future production. In the second case (*"Products not being produced in t"*) the summary statistics are calculated by using just the *future production* dummies of those products that belong to the firm's feasible set *and are not produced by the firm in t.* Formally, the means estimate in this case the *conditional* probabilities of future production.

The first pane (textit "All firms") shows the summary statistics when all the sampled firms are included. The other four panes show the statistics when only the firms that belong to the $q - th$ quartile of non-production workers are included, for $q = 1...4$. This disaggregation across quartiles of non-production workers matters because this variable is used here as a proxy of skilled labour, and the propositions presented in section 3 state that the amount of skilled labour used by a firm determines to some extent its capacity to add new products to its product mix in the future.

The fact that the median is zero for all the cases indicates that most feasible products are not added by the firms to their product mixes, neither in the present time nor in the future (up to 5 years ahead). As expected, the

[13] As explained above, the set of feasible products of a firm f includes all the products that were ever produced by any firm simultaneously with any of the products produced by f in any year.

conditional probabilities are much lower than the unconditional ones. The conditional probabilities truly quantify the probabilities of addition of potential *new* products in the future, as they correspond only to products that are *not* produced in t. On average, between 3% and 7% of the potential new products are added in the future, depending on the time span and the amount of non-production workers. Very importantly, the probability of addition is higher for the highest quartile than for all the other quartiles for all the time spans, and the probability of addition is higher for the second quartile than for the first quartile for all the time spans.

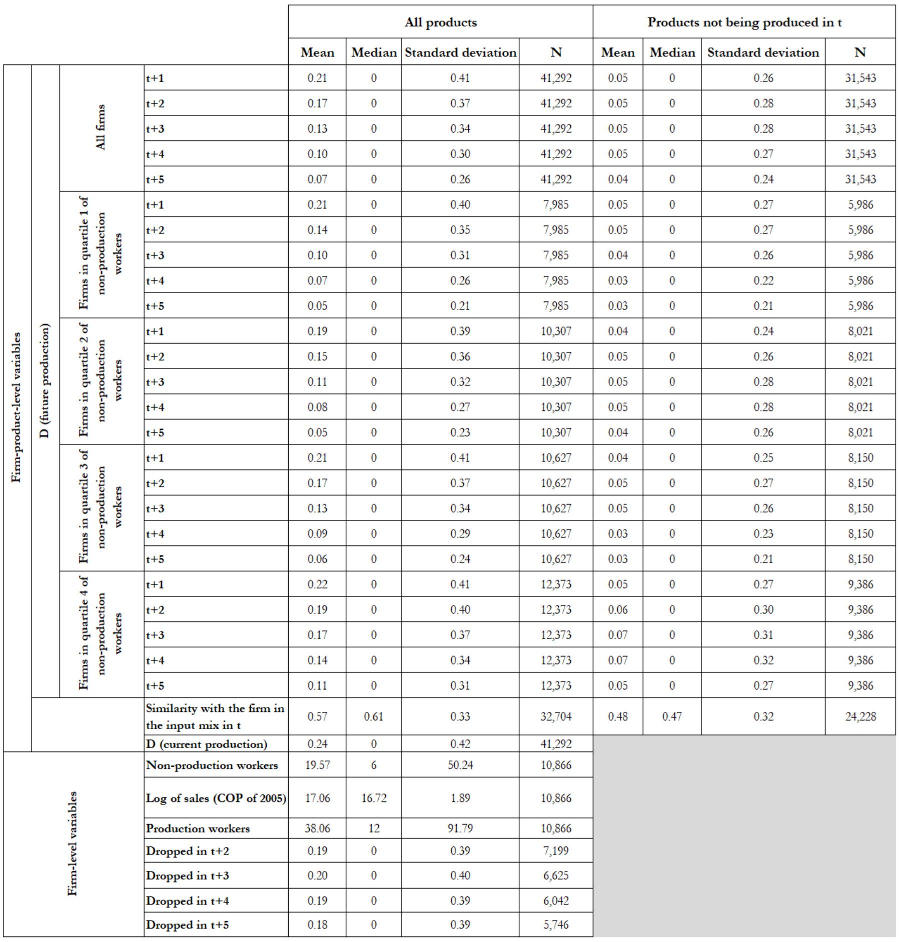
The row below the quartiles (*"Similarity with the firm in the input mix in t"*) shows that the feasible products that are effectively produced by a firm in t are on average more similar to each other than to those feasible products that are *not* produced by this firm in t. In other words, the average similarity of the produced products with the rest of produced products is higher than the average similarity of the non-produced products with the produced products. This is consistent with the propositions presented in Section 2.

As for the statistics for the firm-level variables, it is to note that firms use on average approximately two production workers per each non-production worker. It is also to note that the probability that a firm that exists in t drops at least one product from its product mix in $t+k$ does not change much across different values of k (for $k = 2...5$).

Results

In order to obtain consistent estimators for the parameters in equation (1), I run a regression with ordinary least squares with information from the EAM (see chapter 3 for details of this database). Namely, I use the module of the EAM that includes information of prices and quantities of all the inputs used by all the firms in every year to calculate S_{fpt} for all the products in the feasible set of every firm in the sample. I use the module of the EAM that includes information of prices and quantities of all the products sold by all firms in every year to define the feasible set of each firm and to construct the dummy variables $D_{fp,t+s}$. The number of non-production workers was directly taken from the EAM and used in the regression.

I start by presenting and interpreting the results of estimating the model (1) for $s = 1$, and then I move to longer spans. This because it is convenient to analyze initially the short-term relationship between product addition and the interaction of similarity with non-production workers, and then analyze separately such relation in longer terms. This distinction is relevant in the empirical work presented here because all the regressors (similarity, dummies for quartiles of non-production workers and their interactions) are all defined in the present time t, and their relationships with future product addition might be easier to identify in the short term than in the long term, when other excluded dynamic factors might affect the pattern of product addition to a larger extent than in the short term. I present also the results of regressions in which only the quartiles of non-production workers and the similarity are included separately as regressors.



				All products				Products not being produced in t			
				Mean	Median	Standard deviation	N	Mean	Median	Standard deviation	N
Firm-product-level variables	D (future production)	All firms	t+1	0.21	0	0.41	41,292	0.05	0	0.26	31,543
			t+2	0.17	0	0.37	41,292	0.05	0	0.28	31,543
			t+3	0.13	0	0.34	41,292	0.05	0	0.28	31,543
			t+4	0.10	0	0.30	41,292	0.05	0	0.27	31,543
			t+5	0.07	0	0.26	41,292	0.04	0	0.24	31,543
		Firms in quartile 1 of non-production workers	t+1	0.21	0	0.40	7,985	0.05	0	0.27	5,986
			t+2	0.14	0	0.35	7,985	0.05	0	0.27	5,986
			t+3	0.10	0	0.31	7,985	0.04	0	0.26	5,986
			t+4	0.07	0	0.26	7,985	0.03	0	0.22	5,986
			t+5	0.05	0	0.21	7,985	0.03	0	0.21	5,986
		Firms in quartile 2 of non-production workers	t+1	0.19	0	0.39	10,307	0.04	0	0.24	8,021
			t+2	0.15	0	0.36	10,307	0.05	0	0.26	8,021
			t+3	0.11	0	0.32	10,307	0.05	0	0.28	8,021
			t+4	0.08	0	0.27	10,307	0.05	0	0.28	8,021
			t+5	0.05	0	0.23	10,307	0.04	0	0.26	8,021
		Firms in quartile 3 of non-production workers	t+1	0.21	0	0.41	10,627	0.04	0	0.25	8,150
			t+2	0.17	0	0.37	10,627	0.05	0	0.27	8,150
			t+3	0.13	0	0.34	10,627	0.05	0	0.26	8,150
			t+4	0.09	0	0.29	10,627	0.03	0	0.23	8,150
			t+5	0.06	0	0.24	10,627	0.03	0	0.21	8,150
		Firms in quartile 4 of non-production workers	t+1	0.22	0	0.41	12,373	0.05	0	0.27	9,386
			t+2	0.19	0	0.40	12,373	0.06	0	0.30	9,386
			t+3	0.17	0	0.37	12,373	0.07	0	0.31	9,386
			t+4	0.14	0	0.34	12,373	0.07	0	0.32	9,386
			t+5	0.11	0	0.31	12,373	0.05	0	0.27	9,386
			Similarity with the firm in the input mix in t	0.57	0.61	0.33	32,704	0.48	0.47	0.32	24,228
			D (current production)	0.24	0	0.42	41,292				
Firm-level variables			Non-production workers	19.57	6	50.24	10,866				
			Log of sales (COP of 2005)	17.06	16.72	1.89	10,866				
			Production workers	38.06	12	91.79	10,866				
			Dropped in t+2	0.19	0	0.39	7,199				
			Dropped in t+3	0.20	0	0.40	6,625				
			Dropped in t+4	0.19	0	0.39	6,042				
			Dropped in t+5	0.18	0	0.39	5,746				

Table 1. Summary statistics of relevant firm-product-level and firm-level variables.

Table 1 presents the results of estimating the model (1) for $s = 1$ by ordinary least squares as described before. The third column shows the results when all the regressors in the equation (1) are included. The most important finding is that $\delta^1 + \beta^1_q$ is positive as expected for $q = 2$ and $q = 4$. Moreover, the estimator for the parameter β_4 (that is, for the highest quartile of non-production workers) is larger than the estimators for β_2 and β_3. This means that the correlation between similarity and product addition is unambiguously larger for the firms in the highest quartile non-production workers than for the firms in all the other quartiles. The fact that the estimator for β_2 is positive and statistically significant means that the firms in the second quartile of non-production workers have a higher correlation between product addition and similarity than the firms in the first quartile of this variable.

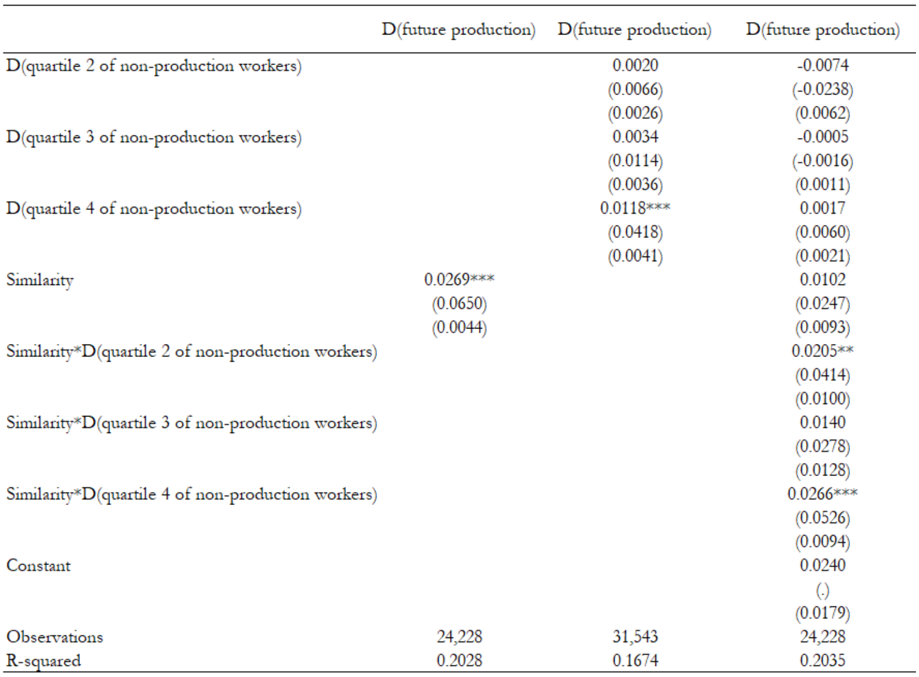
The only fact in Table 2 that contradicts the proposition 1 of section 1 so far is that the correlation between product addition and similarity is not larger for firms in the third quartile than for those in the second quartile. I ran an alternative specification of the model (1) for $s = 1$ in which I included the logarithm of sales as a regressor, in order to avoid a possible omitted variable bias (as the number of non-production workers of a firm is positively correlated with its size, proxied here by its sales). However, the main conclusions summarized so far remain in this case.

Very importantly, the consistency of my empirical results with the theoretical prediction that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$ increases when I include the potential cost of production as a regressor. I will show this result in the section 5 of this paper. The inclusion of this variable might seem redundant with the inclusion of similarity, as it is possible that more similar products are added more easily precisely *because* they are cheaper (and therefore more profitable) for firms. Suppose that a firm f is very good at using glass in t. This makes its cost of producing windows (w) in $t+1$ low. As a consequence, f adds w to its product mix in $t+1$. f is also likely to produce glass-intensive products in t. In this example, the similarity between w and $f's$ production in t in terms of the input mix is high, as both mixes are intensive in glass. In addition, the potential cost of w is low in $t+1$. Both facts (the high similarity of w and its low potential cost) come from the fact that f is good at using glass. Given this, the inclusion of both variables in a regression might be redundant.

However, it is possible that the empirical similarity depends also on other variables different from the drivers of the potential cost. Firms might be forced to use in reality inputs in certain combinations because of physical restrictions in the availability of certain inputs. There may be also technical innovations that lead firms to use temporarily or permanently certain inputs for whose use they do not necessarily have high firm-input-specific productivities. If this were the case, two implications would arise. Firstly, including both the potential cost and the similarity in a regression would not be completely redundant, as they reflect similar but different drivers. Secondly, the similarity might be less correlated than the potential cost with product addition. This because the similarity might indicate just (i) the input mix that a firm *has* to use to produce a product, whereas the potential cost would measure (ii) its intrinsic capacity to produce it cheaply. It is reasonable to expect that (ii) is more correlated with product addition than (i), as long as product addition is more determined by firms' perception of their capacities than by their actual use of inputs.

Given these reasons, I will explore in the section 5 of this paper what happens when both the similarity and the potential cost of production are included. The main conclusion will be that in this case there is stronger evidence that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$, as expected. I interpret this as suggestive evidence of a possible omitted variable bias in the specification (1). Please see the section 5 for a detailed discussion of these results.

As expected, the sum $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ is greater for the fourth quartile than for all the lower quartiles, and $\Gamma_2 + 0.48(\delta^1 + \beta_2^1) > 0.48\delta^1$. However, the same cannot be concluded for the third quartile when compared to the first and second ones. Therefore, the evidence about the hypothesis that firms with more non-production workers are more likely to add new products is not conclusive for $s = 1$.



	D(future production)	D(future production)	D(future production)
D(quartile 2 of non-production workers)		0.0020	-0.0074
		(0.0066)	(-0.0238)
		(0.0026)	(0.0062)
D(quartile 3 of non-production workers)		0.0034	-0.0005
		(0.0114)	(-0.0016)
		(0.0036)	(0.0011)
D(quartile 4 of non-production workers)		0.0118***	0.0017
		(0.0418)	(0.0060)
		(0.0041)	(0.0021)
Similarity	0.0269***		0.0102
	(0.0650)		(0.0247)
	(0.0044)		(0.0093)
Similarity*D(quartile 2 of non-production workers)			0.0205**
			(0.0414)
			(0.0100)
Similarity*D(quartile 3 of non-production workers)			0.0140
			(0.0278)
			(0.0128)
Similarity*D(quartile 4 of non-production workers)			0.0266***
			(0.0526)
			(0.0094)
Constant			0.0240
			(.)
			(0.0179)
Observations	24,228	31,543	24,228
R-squared	0.2028	0.1674	0.2035

Table 2. Results of regression of product addition as a function of similarity in the input mix interacted with dummies for quartiles of non-production workers. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

The numbers below the estimators in Table 1 are the beta coefficients that correspond to each one of them. The beta coefficient for a specific regressor is calculated from its estimator and the standard deviations of both that regressor and the dependent variable. Formally, a beta coefficient for a variable x measures how many standard deviations the dependent variable grows or falls on average when the corresponding regressor grows one standard deviation. This allows an easier economic interpretation of the size of the correlations between the regressors and the dependent variable. The numbers two lines below the estimators are the estimated standard errors. I report beta coefficients and standard errors in this same order in the tables 3 and 4 below. [14].

Summarizing, I have found so far mixed evidence of the possible validity of propositions 1 and 2 of chapter 2. More specifically, the evidence is in favor of these propositions when the firms with the highest values of skilled labour are compared to the rest of firms. Firstly, the firms with the highest values of skilled labour (proxied by the number of non-production workers) exhibit

[14] In the table 5 I report just the p-values because in that case the size of the correlations is not relevant, whereas it will be more important to analyze if the null hypothesis of non-significance of the instrumented variables used to test for causality can or cannot be rejected.

a higher correlation between similarity and product addition than the rest of firms. Similarly, firms in the second quartile of non-production workers exhibit a higher correlation between these two variables than the one observed for the firms in the first quartile. Finally, the expected value of the dummy for product addition one year ahead $D_{fp,t+1}$ is on average higher for the firms in the fourth quartile of the number of non-production workers than for the rest of the firms, as expected.

Let us suppose that there exists a firm $f1$ with a level of skilled labour (non-production workers) that belongs to the second quartile of this variable in t. According to Table 1, an increase in t of one standard deviation of the similarity between a product p and $f1$ in terms of the input mix makes $f1$ 4.1% more likely to add p than a firm in the bottom quartile of skilled labour (non-production workers). In contrast, if other firm $f2$'s skilled labour (number of non-production workers) belongs to the fourth quartile of this variable in t, an increase in t of one standard deviation of the similarity between a product p and $f2$ in terms of the input mix makes $f2$ 5.3% more likely to add p than a firm in the bottom quartile of skilled labour (non-production workers). On the other hand, the average probability of addition for $f1$ of a product with a similarity of (say) 0.4 is $0.024 + 0.0102 * 0.4 + 0.0205 * 0.4 = 3.6\%$. In contrast, the average probability of addition for $f2$ of a product with a similarity of (say) 0.4 is $0.024 + 0.0102 * 0.4 + 0.0266 * 0.4 = 3.9\%$.

Given that the results found for $s = 1$ are partially in favor of the propositions 1 and 2 of section 2, it is reasonable to expect the same for $s > 1$. In order to analyze the possible empirical validity of these propositions for $s > 1$, I ran four regressions with ordinary least squares that correspond to the model (1) for $s = 2...5$. Table 3 shows the results.

A first fact that stands out from Table 3 is that the number of observations falls as the time span becomes longer in most of cases. This happens because of two factors. Firstly, some firms in my sample disappear as the time passes, so the number of firms that remain in the sample for s years falls as s increases (that is, as the span of analysis becomes longer). Secondly, the number of years included in the regressions falls as s becomes larger, as in any regression that includes the lead of at least one variable. Namely, one year is lost every time s increases one unit (year). As the total number of observations in the regression for a specific s equals the total number of years minus s multiplied by the number of firms that survive s years after each year in the sample multiplied by the number of products in the feasible set of each firm (which does not change over time for a firm), the observed decreases in the total number of observations is exclusively attributable to the two factors mentioned above.

Another fact that must be mentioned about the regressions in Table 3 is that the fitting power of the models is larger for longer spans than for $s = 1$. The coefficient of determination ($R2$) is greater than 30 percent in three of the four cases, well above the $R2$ of 20% for $s = 1$. This might be due to several factors. Firstly, the inclusion of the variable *"Dropped product(s) in t+s"* might be increasing the proportion of total variability in product addition predicted by the model. This variable will be explained in detail later in this section. On the other hand, $R2$ might also be higher because the firms included in the regressions exhibit less variability in their patterns of product addition as s grows from 1 to higher numbers.

	D(future production in t+2)	D(future production in t+3)	D(future production in t+4)	D(future production in t+5)
D(quartile 2 of non-production workers)	0.0074	0.0193***	0.0243***	0.0298***
	(0.0224)	(0.0590)	(0.0781)	(0.1022)
	(0.0069)	(0.0068)	(0.0073)	(0.0101)
D(quartile 3 of non-production workers)	0.0041	0.0111	0.0106	0.0133
	(0.0128)	(0.0344)	(0.0353)	(0.0477)
	(0.0076)	(0.0075)	(0.0077)	(0.0094)
D(quartile 4 of non-production workers)	0.0076	0.0151*	0.0176*	0.0141
	(0.0256)	(0.0516)	(0.0640)	(0.0553)
	(0.0080)	(0.0091)	(0.0092)	(0.0104)
Similarity	0.0175*	0.0329***	0.0301***	0.0226*
	(0.0395)	(0.0744)	(0.0720)	(0.0583)
	(0.0102)	(0.0108)	(0.0108)	(0.0134)
Similarity*D(quartile 2 of non-production workers)	0.0086	-0.0242**	-0.0258**	-0.0244*
	(0.0158)	(-0.0461)	(-0.0504)	(-0.0509)
	(0.0111)	(0.0118)	(0.0115)	(0.0136)
Similarity*D(quartile 3 of non-production workers)	0.0045	-0.0259**	-0.0182	-0.0143
	(0.0082)	(-0.0467)	(-0.0350)	(-0.0297)
	(0.0130)	(0.0128)	(0.0120)	(0.0141)
Similarity*D(quartile 4 of non-production workers)	0.0169	-0.0166	-0.0092	-0.0063
	(0.0326)	(-0.0318)	(-0.0184)	(-0.0136)
	(0.0123)	(0.0126)	(0.0127)	(0.0146)
dropped	0.0220***	0.0201***	0.0226***	0.0142***
	(0.0586)	(0.0563)	(0.0660)	(0.0440)
	(0.0044)	(0.0039)	(0.0041)	(0.0034)
Constant	0.0335	0.0606*	0.0895**	0.1042***
	(.)	(.)	(.)	(.)
	(0.0287)	(0.0318)	(0.0360)	(0.0365)
Observations	17,622	16,218	14,737	14,048
R-squared	0.2932	0.3021	0.3114	0.3096

Table 3. Results of regressions of product addition in future years from $t + 2$ to $t + 5$ as a function of similarity in the input mix interacted with dummies for quartiles of non-production workers. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

As expected, δ^s is positive and statistically significant for every $s = 2...5$. Also as expected, $\delta^s + \beta_q^s$ is positive for all the non-excluded quartiles of non-production workers (that is, for all q from 2 to 4) and for every $s = 2...5$, with the only exception of the firms in the lowest quartile of non-production workers in $t + 5$. These facts imply that the derivative of $D_{fp,t+s}$ with respect to similarity is positive for all quartiles and spans, except for the firms in the lowest quartile of non-production workers in $t + 5$. However, I cannot conclude that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$ for any s. Therefore, the aforementioned derivative does not grow as I move to higher quartiles of non-production workers.

The evidence does not allow me to conclude in general that the sum $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ grows as q grows. In other words, there is not evidence that the firms with more non-production workers are more likely to add new potential products for $s = 2...5$.

As for the evolution of estimators over time, I cannot conclude from Table 3 that δ^s and $\beta_q^s + \delta^s$ fall over time in all cases, as expected. The evidence is mixed in this case. To start, δ^s grows from $s = 2$ to $s = 3$, but then falls monotonically from $s = 3$ to $s = 5$. As for $\beta_q^s + \delta^s$, it falls monotonically from $s = 2$ to $s = 5$ for quartile 2. For quartiles 3 and 4 it falls from $s = 2$ to $s = 3$, then grows from $s = 3$ to $s = 4$, and falls again from $k = 4$ to $k = 5$.

Summarizing, I found that the derivative of product addition with respect to similarity is positive on average for firms of all the quartiles of non-production workers in all spans from $s = 2$ to $s = 5$. This is consistent with the expected results. However, I did not find conclusive evidence that this derivative is in general greater in firms with more non-production workers. As for the dynamics of this derivative, I could not establish conclusive evidence that it falls over time, as expected.

The variable ”Dropped” is a dummy equal to one if the firm drops a product in the respective year $t + s$, and zero otherwise. It intends to capture the possibility (not explored in the model of the second chapter of this thesis) that firms has scarce factors of production that are necessary for the production of several products and that cannot be used simultaneously for the production of several products to a full extent, and that therefore they need to drop some existing products to add others. The possible existence of these factors was theorized by Sutton (2012), who considers it a key determinant of product scope of countries. This variable is positive and statistically significant at a level of significance of 1 % for every $s = 2...5$, which suggests the possible presence of factors of production with the properties mentioned above. A deep exploration of their role and nature at the firm level remains as a pending and crucial task for future works.

It is worth to mention that the analyses of significance presented so far might be influenced by the number of firms that I am using. Namely, my sampling strategy implies that I am using only 160 firms, which in turn implies that I use just 160 clusters to estimate the standard errors of the estimators. A low number of clusters may lead in general to incorrect rejections of null hypotheses that true parameters are equal to zero[15] (this is, this can lead to claiming statistical significance in cases in which it does not exist). However, 160 is a larger number than the minimum number of clusters identified in the literature for the statistical inference to be still valid (which is about 50) [16].

Robustness check: keeping the same sample of firms across years

It is possible that the changes of estimators across different spans for a given quartile of non-production workers described in the previous section be partially attributable to changes in the sample. As I explained before, some firms disappear from the sample as I move to longer spans (higher values of s) and the sample size falls mechanically as I do so. These facts might explain to some extent that changes of estimators across different spans differ from the expected ones. In order to explore this possibility, Table 4 shows results for the same regressions of Table 3, with the only difference that the regressions in table 4 are all run with the same sample. Namely, I ran all the regressions of Table 4 using only the firms that survived continuously from t to $t+5$ and using in every case only information of the years that can be used for all the four regressions.

The main conclusion from Table 4 is that the results after restricting the sample of firms to avoid a possible attrition bias are not very different from the ones that I found with the whole sample of firms. In this restricted case, $\delta^s > 0$ and $\delta^s + \beta^s_q > 0$ for $s = 3$, $s = 4$ and $s = 5$, and for all quartiles of non-production workers. In words, the derivative of $D_{fp,t+s}$ with respect to similarity is positive for all quartiles for all those three spans (3, 4 and 5 years ahead). This is not very different from the results in Table 3. Just as for Table 3, I cannot conclude from table 4 that $0 \leq \beta^s_2 \leq \beta^s_3 \leq \beta^s_4$ for any s. Therefore, the aforementioned derivative does not grow either in this case as I move to higher quartiles of non-production workers. Once again, I cannot conclude from Table 4 that $\Gamma_q + 0.48(\delta^s + \beta^s_q)$ grows as q grows in general for $s = 2...5$. Finally, I cannot conclude from Table 4 that δ^s and $\beta^s_q + \delta^s$ fall over time in all cases, as expected. This conclusion is identical to the one I presented for the case with the whole sample of firms.

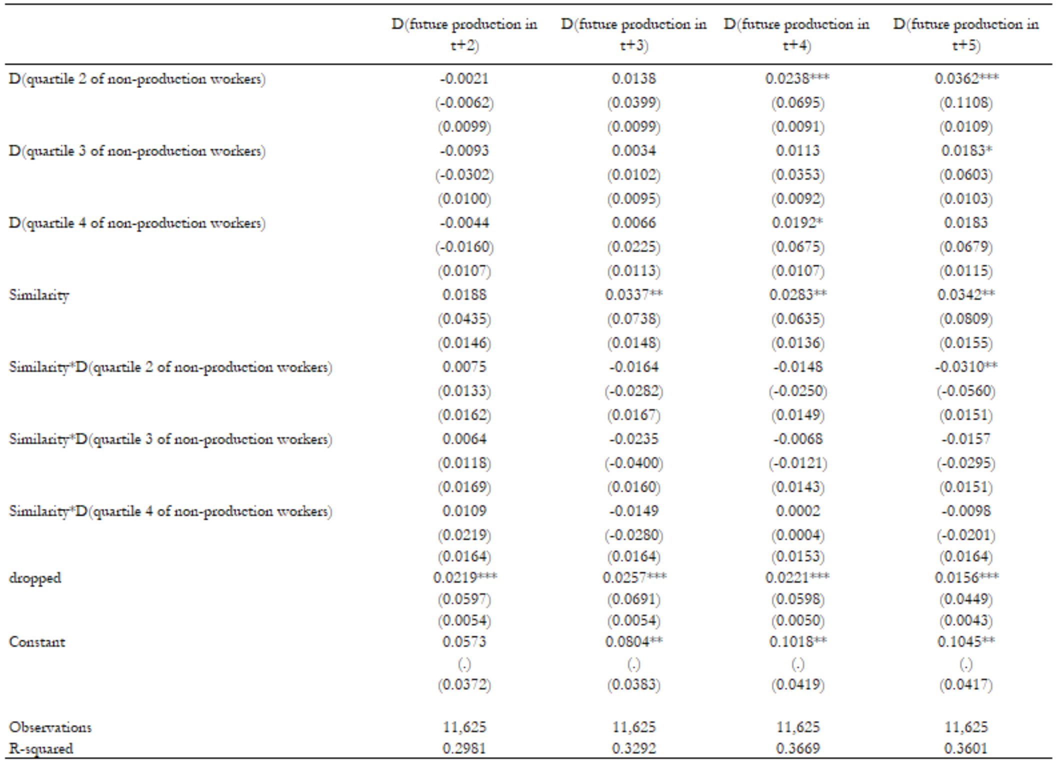
In addition to the possible problems induced by the exit of firms over time, I had the concern that the variability of the dependent variables of the regressions whose results are shown in tables 2 and 3 might be too low. This because the dummies for future product addition have too many zeros, as the feasible sets include many more products than the ones that are in fact produced by the firms. This problem might have two negative implications. Firstly, it might reduce the statistical power of the models estimated here. Secondly, the variance of the errors might be too low, leading to incorrect inferences about the statistical significance of the estimators. In order to analyze if this is the case, I ran the regressions of table 3 with a more restrictive definition of the feasible set.

I defined the feasible set in an alternative way in which this set contains for a firm f all the products that f has ever produced and all the products with

[15] Please read Cameron et al. (2008) for details.

[16] Please read Angrist and Pischke (2009) for a detailed explanation.

a probability of joint production with any product ever produced by f above the median of the distribution of all the probabilities of joint production (unlike the original definition, in which this set includes *all* the products that were ever jointly produced with any of the products ever produced by f). This reduces the number of zeros in the dependent variable (the dummy for product addition), and prevents the problem of too small variability explained before. However, I did not observe important changes in this case with respect to the results shown in table 3.



	D(future production in t+2)	D(future production in t+3)	D(future production in t+4)	D(future production in t+5)
D(quartile 2 of non-production workers)	-0.0021	0.0138	0.0238***	0.0362***
	(-0.0062)	(0.0399)	(0.0695)	(0.1108)
	(0.0099)	(0.0099)	(0.0091)	(0.0109)
D(quartile 3 of non-production workers)	-0.0093	0.0034	0.0113	0.0183*
	(-0.0302)	(0.0102)	(0.0353)	(0.0603)
	(0.0100)	(0.0095)	(0.0092)	(0.0103)
D(quartile 4 of non-production workers)	-0.0044	0.0066	0.0192*	0.0183
	(-0.0160)	(0.0225)	(0.0675)	(0.0679)
	(0.0107)	(0.0113)	(0.0107)	(0.0115)
Similarity	0.0188	0.0337**	0.0283**	0.0342**
	(0.0435)	(0.0738)	(0.0635)	(0.0809)
	(0.0146)	(0.0148)	(0.0136)	(0.0155)
Similarity*D(quartile 2 of non-production workers)	0.0075	-0.0164	-0.0148	-0.0310**
	(0.0133)	(-0.0282)	(-0.0250)	(-0.0560)
	(0.0162)	(0.0167)	(0.0149)	(0.0151)
Similarity*D(quartile 3 of non-production workers)	0.0064	-0.0235	-0.0068	-0.0157
	(0.0118)	(-0.0400)	(-0.0121)	(-0.0295)
	(0.0169)	(0.0160)	(0.0143)	(0.0151)
Similarity*D(quartile 4 of non-production workers)	0.0109	-0.0149	0.0002	-0.0098
	(0.0219)	(-0.0280)	(0.0004)	(-0.0201)
	(0.0164)	(0.0164)	(0.0153)	(0.0164)
dropped	0.0219***	0.0257***	0.0221***	0.0156***
	(0.0597)	(0.0691)	(0.0598)	(0.0449)
	(0.0054)	(0.0054)	(0.0050)	(0.0043)
Constant	0.0573	0.0804**	0.1018**	0.1045**
	(.)	(.)	(.)	(.)
	(0.0372)	(0.0383)	(0.0419)	(0.0417)
Observations	11,625	11,625	11,625	11,625
R-squared	0.2981	0.3292	0.3669	0.3601

Table 4. Results of regressions of product addition in future years from $t + 2$ to $t + 5$ as a function of similarity in the input mix interacted with dummies of quartiles for non-production workers keeping the same sample. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

Discussion

The results presented in this section are mostly indicative of the existence of a positive derivative of product addition with respect to similarity that remains over time, although it does not change over time in the way I expected. In simpler words, firms keep on adding to their product mixes even in the long term products whose production required in t input mixes that were more similar to the input mixes used by them in that year. This finding is important because this persistence is consistent with the main and distinctive feature of the model that I proposed in the second chapter of this thesis, which is the persistence of firm-input-specific productivities that cause persistent patterns of product addition.

On the other hand, the empirical results presented here are not fully consistent with the expected result that the derivative of product addition with respect to similarity is greater in firms with more non-production workers. However, I did find that the firms with the highest values of non-production workers exhibit a larger derivative than the rest of firms in the short term (in $t + 1$).

The results presented here do not suggest unambiguously that the derivative of product addition with respect to similarity falls over time for every quartile of non-production workers, as expected. There are several possible explanations for this. One that is consistent with the model of the second chapter of this thesis is that firm-input-specific productivities might depreciate over time. If this were the case, such productivities might grow or fall over time, depending on the size of the depreciation and on the scale of the process of learning by using. Depreciation might dominate in some periods, leading firms to add more (and not less) similar products, as they must rely to a larger extent in this case on what they already know how to do well. I do not explore this or other possible explanations in this paper.

5 The effect of cheaper inputs on product addition

In this section I analyze if the exogenous changes in the prices of inputs have in reality the firm-product-specific effects predicted by the theoretical model presented in the second chapter of this thesis. Very importantly, I explain here why the granularity of this prediction allows me to claim that finding possible validity for it would constitute causal evidence in favor of the theoretical model presented in the second chapter of this thesis.

The Proposition 3 in section 2 predicts what is expected to happen to product addition if the price of an input used by the firms falls. It states that this reduction has firm-product-specific effects on the profitability of product addition. More specifically, this effect depends in each case on two differential elements: (i) it is larger if the respective product is intensive in the input whose price falls, and (ii) it is larger if the respective firm has higher firm-input-specific productivity for the use of the inputs whose price falls.

The ideal way to analyze empirically the possible validity of the firm-product-specific prediction contained in the proposition 3 would be to construct a variable that reflects the two elements (i) and (ii) mentioned above, and analyze its possible correlation with the phenomenon of product addition. This variable should ideally have two characteristics. Firstly, it should be possible to calculate it in reality, as its observed values are needed if I want to use it in an empirical analysis. Secondly, it should *capture* the two differential elements (i) and (ii) above. By "capture" I mean that this variable should respond to the changes in the prices of inputs in directions and magnitudes that reflect the elements (i) and (ii) above.

Fortunately, the theoretical model in the second chapter of this thesis offers an ideal candidate for the variable mentioned above. The expression (6) of the second chapter of this thesis is the conditional cost function of a firm f to produce a quantity Q_{fp} of a product p in a year t. Formally, this cost has the following functional form:

$$\mathrm{C}_{fpt} = \left[\sum_k \phi^{\sigma}_{fkt}\phi^{\sigma}_{pk}q^{1-\sigma}_{fkt}\right]^{1/1-\sigma} \frac{Q^{1/\theta}_{fp}}{g_p\left(\overline{K_{ft}},\overline{L_{ft}}\right)^{\beta}/\theta}$$

(3)

C_{fpt} has the key property that it determines the phenomenon of product addition (this is why it is the first result presented in the second chapter of this thesis). Intuitively, if a firm f can produce a product p in a year t at a lower cost (this is, if C_{fpt} above is lower), then p is more profitable for f and it is easier for f to add it to its product mix in the future. In short, if C_{fpt} is higher, then the profitability of adding p for f in t is lower.

Very importantly, C_{fpt} is lower *for every produced quantity* if the term $\left[\sum_k \phi^{\sigma_p}_{fkt}\phi^{\sigma_p}_{pk}q^{1-\sigma_p}_{fkt}\right]^{1/1-\sigma_p}$ is lower. In other words, this component scales up or down the minimum cost of production for any produced quantity. Given its relevance for the analysis below, I will name this term *CES cost index* (CCI) hereinafter, as I will use it repeatedly later. Formally:

$$\mathrm{CCI}_{fpt} \equiv \left[\sum_k \phi^{\sigma}_{fkt}\phi^{\sigma}_{pk}q^{1-\sigma}_{fkt}\right]^{1/1-\sigma}$$

(4)

What makes CCI_{fpt} special for the purposes of this paper is that it captures formally the intuition explained before for the facts that the firm-product-specific effect of the reduction in the price of an input is (i) larger if the respective product is intensive in the input whose price falls, and (ii) larger if the respective firm has a higher firm-input-specific productivity for the use of the input whose price falls. In other words, CCI_{fpt} has the desired property of capturing (i) and (ii). The part (i) is captured by the fact that the price of each input q_{fkt} is multiplied by ϕ_{pkt} (the input-output coefficient), which amplifies the effect of falls in prices of inputs with high ϕ_{pk}. The part (ii) is captured by the fact that the price of each input q_{fkt} is also multiplied by ϕ_{fkt} (the firm-input-specific productivity), which amplifies the effect of falls in prices of inputs with high ϕ_{fkt}. Please notice that the fact that both effects are explained by the triple product $\phi_{fkt}\phi_{pk}q_{fkt}$ (each with its respective exponential) implies that the two differential factors (i) and (ii) amplify each other. This latter fact is not explored theoretically nor empirically in this thesis, but it might be explored in the future.

So far, I have explained why CCI_{fpt} has the key characteristic of capturing properly (i) and (ii). In short, CCI_{fpt} determines the cost of production (which in turn determines the phenomenon of product addition), and it changes when the prices of inputs fall in firm-product-specific magnitudes that reflect the differential elements (i) and (ii). However, I have not demonstrated yet that CCI_{fpt} has the other desired property of being possible to calculate. Fortunately, CCI_{fpt} can be calculated indeed, as I will show later in this section. Before explaining how, I will explain how it is used, what are the expected results when using it, and what econometric problems might arise because of its use.

Econometric model and expected results

Given the very important characteristics of CCI_{fpt}, I use it as a regressor for the probability of product addition. More specifically, I include it as an additional regressor in the first econometric model of the previous section. Namely, I estimate here the parameters of the following expression:

$$\mathrm{D}_{fp,t+1} = \alpha + \textstyle\sum_{q=1}^{4} \beta_q^s S_{fpt} * D_{qft} + \sum_{q=1}^{4} \Gamma_q^s D_{qft} + \Gamma log(CCI_{fpt}) + \delta S_{fpt} + \psi_t + \phi_f + \tau_p + \epsilon_{fp,t+1} \quad (5)$$

All the interpretations and explanations of expected signs provided for the parameters in the expression (1) apply entirely for the analogous parameters in the expression (5). The expected sign for Γ is negative, as a lower CCI_{fpt} should reduce the cost of production of p for f in t and this should in turn affect positively the potential profitability.

Very importantly, CCI_{fpt} changes when the prices of inputs change in such a way that it captures in theory the key differential elements (i) and (ii) explained in detail above. As these very granular and specific elements are predicted by the theoretical model presented in the second chapter of this thesis in its more specific and granular proposition, their possible empirical validity when the prices of inputs change exogenously may be interpreted as *causal* evidence in favor of this model.

Estimation of structural parameters

As I mentioned above, it is possible to compute CCI_{fpt}. For this, let us start by writing the first order condition of the cost equation of a firm f to produce a product p in a year t with respect to the material s (this is, with respect to M_{fpst}). As I showed in the expression (17) of the second chapter of this thesis, this first-order condition is as follows[17]:

$$g_p\left(K_{ft}, L_{ft}, H_{ft}\right)^{\beta_p(1-\eta_p)} \frac{\theta_p(1-\eta_p)}{\rho_p} \left[\sum_k \phi_{fkt}\phi_{pk}M_{fpkt}^{\rho_p}\right]^{\frac{\theta_p(1-\eta_p)}{\rho_p}-1} \phi_{fst}\phi_{ps}\rho_p M_{fpst}^{\rho_p-1} = q_{fst}$$
(6)

It is possible to identify ϕ_{fst} and ϕ_{ps} from this expression with a combination of some simple transformations and OLS regressions. The expression (6) can be rewritten in terms of the observable revenue perceived by the firm from sales of product p in t (represented by R_{fpt}) as follows:

$$\frac{R_{fpt}}{q_{fst}} = \frac{\alpha_{fpt}}{\theta_p(1-\eta_p)\phi_{fst}\phi_{ps}M_{fpst}^{\rho_p-1}}$$
(7)

where $\alpha_{fpt} = \left[\sum_k \phi_{fkt}\phi_{pk}M_{fpkt}^{\rho_p}\right]$. The ratio in the left-hand side of (7) is observable because both the revenues from selling all the products and the prices of the inputs are observables for every firm. I represent such ratio here by y_{fpst}. In addition, I take the natural logarithm of both sides of the equality in (7) to get the following expression:

$$lny_{fpst} = \psi_{fpt} - \delta_{ps} - \Omega_{fst} - \epsilon_{fpst}$$
(8)

where $\psi_{fpt} = log\left[\frac{\alpha_{fpt}}{\theta_p(1-\eta_p)}\right]$, $\delta_{ps} = log\phi_{ps}$, $\Omega_{fst} = log\phi_{fst}$ and $\epsilon_{fpst} = logM_{fpst}^{\rho_p-1}$. The important feature of expression (8) is that all its unknown components can be estimated by running an OLS regression of the observables

[17] Please see the second chapter of this thesis for an explanation of all the parameters and variables involved in the expression (6)

lny_{fpst} as a function of firm-product-year, product-input and firm-input-year fixed effects (which will account for ψ_{fpt}, δ_{ps} and Ω_{fst}, respectively). This is important because by estimating δ_{ps} and Ω_{fst} it is possible to recover two key components of CCI_{fpt}: the firm-input-specific productivities ϕ_{fst} and the input-output coefficients ϕ_{ps}.

I use the OLS estimators of ϕ_{fst} and ϕ_{ps} in this chapter to calculate CCI_{fpt}. In addition, they are used in the section 6 of this paper to analyze possible empirical evidence of the existence of the main mechanism and main assumption of the model presented in the second chapter of this thesis.

Given that I have now estimators for ϕ_{ps} and ϕ_{fst} and the prices paid by all the firms for all the inputs are observables, all I still need to calculate CCI_{fpt} is a value for σ. I take its value from the work from Eslava and Haltiwanger (2020), who also used data from the Colombian EAM and performed a GMM estimation of the elasticity of substitution across materials of Colombian manufacturing firms. They assumed a CES functional form very similar to the one I use here. Because of this comparability in terms of the used data and the functional form, their value is highly applicable to this paper. Their average across all sectors of these elasticities of substitution is 1.84. This is the value that I use here to calculate CCI_{fpt}.

Once calculated, CCI_{fpt} is used in the regression presented in the expression (5) above. Very importantly, CCI_{fpt} can be estimated for the products not produced by the firms, and not only for the products that they actually produce. This is crucial, as I use it in this section in a regression with a dummy for product addition in $t+1$ of products *not produced in* t as the dependent variable. Therefore, it will have to be used by definition for products that are not produced in t.

Possible endogeneity

There may be a problem of endogeneity of the variable CCI_{fpt} in the expression (5). Namely, this variable might be correlated with the variables included in $\epsilon_{fp,t+1}$. For instance, the price that the sellers of inputs charge a firm f for the inputs needed to produce a product p can depend on their own forecast about the future decision by f to produce or not p in $t+1$. If the seller of the inputs forecasts that the firm will not produce p in $t+1$, it might reduce the prices of inputs in t (especially for perishable inputs). If this forecast is based on variables not included in the model (which is likely), this would mean that CCI_{fpt} would be correlated with $\epsilon_{fp,t+1}$, and the OLS estimator of Γ would be biased. To solve this, I perform here a two-stage least squares (2SLS) estimation, using tariffs as instruments.

The Colombian trade reform of 2012

Colombia is a small open economy [18]. This means that it is reasonable

[18] The term "open" means in this context that goods and services can be traded from and to the country, even though some of them are subject to tariffs.

to assume that it is a price taker in the tradable products sold under perfect competition, and that prices of imports should have influence on the prices of the tradable products sold under monopolistic competition. Therefore, if the imports of the tradable inputs used by a firm f become cheaper, the prices of tradable inputs paid by f should fall, even if f buys them from domestic producers. Formally, this means that there should be a positive relationship between the price paid by f for an input k in t (q_{fkt}) and the tariff imposed by the domestic government to input k in t (T_{kt}).

A simple way to justify formally the assumption that q_{fkt} is positively correlated with T_{kt} comes from the optimal pricing behavior of firms under free trade. In general, the price charged by sellers of domestic inputs to manufacturing firms is equal to the marginal cost of the respective input multiplied by a factor F. This factor can depend on several other factors, such as the elasticity of substitution and the number of firms. It is constant under most of the models typically used in economics, and it is equal to one under perfect competition.

Formally, for domestic inputs $q_{fkt} = F.MC_{fkt}$, where MC_{fkt} is the marginal cost of producing the input k faced by the domestic firm that produces it and sells it to f in t. Analogously, the price of imported inputs is $q^*_{fkt} = F.MC^*_{fkt}.(1 + T_{kt})$, where MC^*_{fkt} is the marginal cost of producing the input k faced by the foreign firm that produces it and sells it to f in t. If the tariff T_{kt} decreases, q^*_{fkt} also decreases. Unless the domestic producer of an input k is a monopolist, its price q_{fkt} falls when q^*_{fkt} falls, because F falls for k as a consequence of the higher competition. Under perfect competition, the fall in the tariff is fully transmitted to the price.

Given this, the policy action used here as a source of exogenous variation in the prices of inputs is a unilateral reduction of the tariffs imposed by the Colombian government to the imports from the United States that took place in 2011, prior to the free-trade agreement between these two countries in 2013. As figure 1 shows, the simple average of the product-level tariffs imposed by the Colombian governments to the imports from the United States fell almost 3 percentage points from 2010 to 2011, and kept on falling during almost all the successive years. In eighteen years this average fell almost 8 percentage points, from near 12 percent in 2000 to approximately 4 percent in 2018.

The figure 2 shows that it was not only the average of the product-level tariffs to the imports from United States that changed, but also the distribution of such tariffs. The upper and lower limits of blue boxes in this figure represent the percentiles 25th and 75th of the distribution of tariffs each year, respectively. The horizontal line inside each box is the median of each year. The two horizontal lines above and below each box are the upper adjacent values and the lower adjacent values each year, respectively [19].

Figure 2 shows several facts. The most important is that the distribution

[19] The upper adjacent value is uniquely defined as the j percentile of the distribution such that complies with two requirements: (A) it is smaller than percentile(75)+1.5*(percentile(75)-percentile(25)), and (B) the j+1 percentile is larger than this value. The lower adjacent value is defined in an analogous way

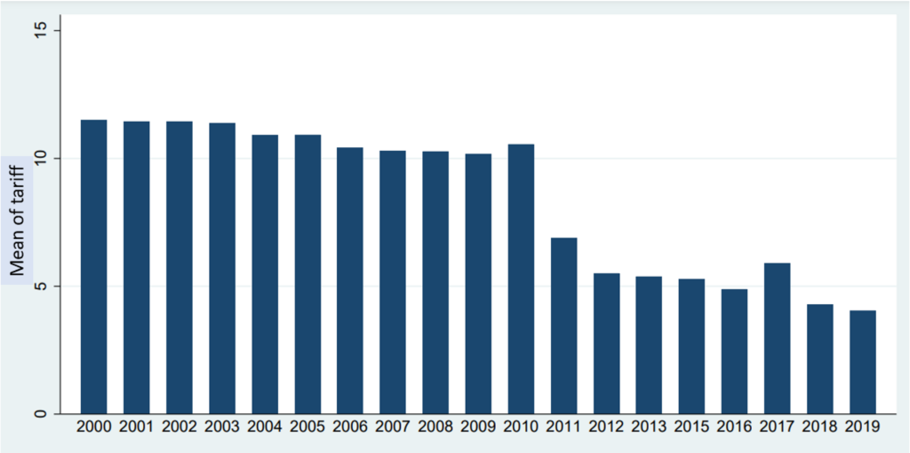


Figure 1. Average of product-level tariffs imposed by the Colombian government to the imports from the United States.

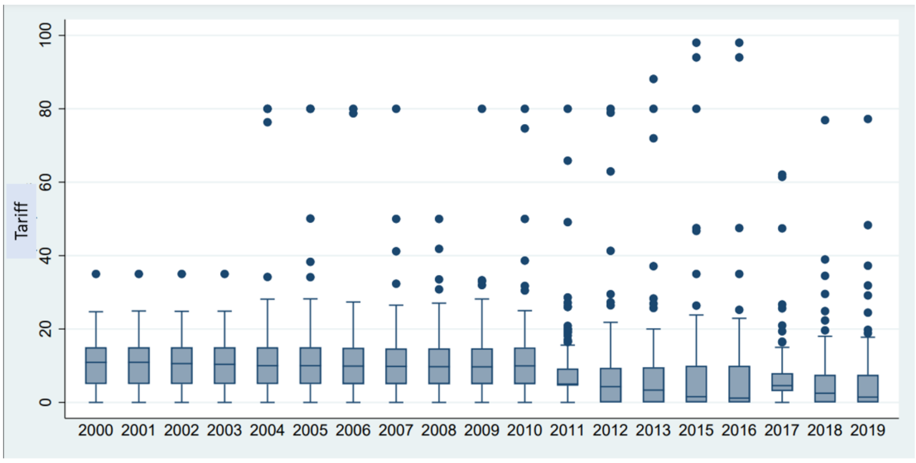


Figure 2. Box plots of yearly distributions of product-level tariffs imposed by the Colombian governments to imports from the United States.

of Colombian tariffs to imports from the United States changed in many senses in 2011. Namely, the interquartile range (75th percentile - 25th percentile) fell from nearly 10 percentage points in all years from 2000 to 2010 to approximately 5 percentage points in 2011. In this last year 50% of the tariffs were between 5 and 10%, which moved the median down from 10% in previous years to 5% in 2011. Several values that would have been below the upper adjacent values in previous years became outliers in 2011. This change in the concentration of the distribution towards lower values continued in the successive years. As a consequence, the median of the distribution fell almost to zero in 2016. Even though it increased temporarily to nearly 3 percent in 2017, it fell again since then until 2019, when it reached again a value very close to zero.

The fact that the median of product-level tariffs imposed by the Colombian government to the imports from the United States fell nearly 5 percentage points from 2010 to 2011 means that many tariffs changed, and that the reduction was not applied just to a few products. This conclusion is reinforced by the other changes in the distribution that can be observed in Figure 2 from 2010 to 2011. The 25th percentile seems to have remained the same (around 5 percent), but the mean fell from nearly 10 percent to nearly 5%. These facts mean that the second quartile of the distribution squeezed. More specifically, 25% of the products had until 2010 a tariff between 5 and 10%, and they (or other products that replaced them in the second quartile) switched to have a tariff of 5% or below.

As a consequence of the squeezing of the lower quartiles described before and of possible reductions of tariffs in the upper parts of the distribution, the 75th percentile also fell nearly 5 percentage points from 2010 to 2011, from nearly 15 percent to 10%. In summary, the change in tariffs that took place in 2011 was not a reduction in just a few products. Instead, it affected a proportion of the products sufficiently high as to cause the notorious change in the distribution that can be seen in Figure 2. The same is true for the years after 2011. Additional checks confirm that the tariffs of more than half of the products fell in 2011, and the same is true for the subsequent years.

The fact that the reduction in tariffs from 2011 was broad both in its extensive (number of affected products) and intensive (extent of changes) margins allows me to use it as a phenomenon that affected most of the products used as inputs by the Colombian manufacturing firms. As for the exogeneity of this phenomenon, it is reasonable to assume that the residuals ϵ_{fpt} in expression (1) in 2011 and in the subsequent years were uncorrelated to the decision of the government to reduce the tariffs of the inputs that are critical for the production of p by firm f since 2011.

Even though some factors that might affect product addition by a firm f such as its capacity of agency or lobbying might be correlated to the decision of the government to reduce or not the tariffs of the inputs used by f, I discard the possibility that this is a generalized fact. This because the median of the market share of a firm in the total market of a product in a year is below 30 percent, even though the definition of product used here is as narrow as possible (that is, I use the most disaggregated level of product definition available in the different product classifications used here).

Results

If the generalized reduction in tariffs that occurred from 2011 led to a fall in the prices of the inputs used by the Colombian manufacturing firms, this generalized reduction should cause a decrease of the right-hand-side of expression (4). Given this relevance as instruments, I use here the average tariffs imposed by the Colombian government to imports from the rest of the world as instruments for CCI_{fpt} in a 2SLS regression.

In the first stage I regress CCI_{fpt} on the two principal components of the tariffs in t on all the inputs used by the firm in t. These two principal components are uncorrelated with each other and they have two key properties: (i) they capture the maximum possible variance of the original tariffs, which means that they capture to a large extent the most notorious differences across tariffs within a firm in a year, and (ii) they are therefore a two-dimensional comprised version (in terms of variance) of the original tariffs. Therefore, they capture to the largest possible extent the variability of these latter while still allowing me to have just two uncorrelated variables derived from *all* the tariffs for each firm.

The two-dimensional nature of principal components allows me to run a unique regression for all the firm-year-product combinations in the first stage. This regression has just two explanatory variables in all cases, unlike the potential situation in which I had used all the relevant tariffs for each firm in each year[20]. In the second stage I used the predicted values from the first stage $CC\hat{I}_{fpt}$ as an explanatory variable in (5).

Table 5 shows the results both using such principal components as instruments for CCI_{fpt} and using CCI_{fpt} itself as a non-instrumented regressor.

I report results in all the specifications included in the table (5) both including and excluding similarity and its interactions with dummies for quartiles of non-production workers. This because there are valid reasons both to include them and to exclude them.

The similarity might be highly correlated with CCI_{fpt}. If a firm f is very proficient in the use of a particular input k in t and the product p is very intensive in k, then CCI_{fpt} will be low. If these assumptions hold, it is likely that f uses k intensively, and also that the production of p is in general intensive in k. In other words, it is likely that the similarity between f and p in the use of inputs is high, which in mathematical terms would mean that S_{fpt} is high. Therefore, there are theoretical reasons to expect a high negative correlation be-

[20] It would be possible in theory to run firm-year-specific regressions of CCI_{fpt} as a function of all the tariffs relevant for the firm in t. However, the number of observations for each firm in this case would be equal to the number of produced actually products by it in t. As each manufacturing firm produces on average less than four products in a year, the number of observations of each regression in this second specification would be very low in many cases. This problem can be reinforced by the fact that some firms use many inputs, which increases the number of parameters to be estimated.

tween CCI_{fpt} and S_{fpt}. The negative correlation between these two variables is indeed slightly above -0.5 in absolute value. Therefore, my estimator for Γ might be inconsistent if I exclude S_{fpt} and its interactions from (5).

However, there are also theoretically valid reasons to exclude S_{fpt} and its interactions from (5), as it is possible to argue that CCI_{fpt} and S_{fpt} are conceptually redundant and they do not need and should not be included in the same regression, as the same fundamentals drive them both. If a firm f is very good at using an input k and some products are very intensive in k, the model presented in the second chapter of this thesis predicts that such products are more likely to be added by f than others because f can produce them efficiently (this is, because CCI_{fpt} is low). This higher relative efficiency *is materialized in reality* by the fact that f uses k intensively, just like k is intensively used to produce those products. This similar intensity implies a high S_{fpt}. In short, similarity is just a metric that *reflects and materializes efficiency*, and its exclusion should not generate any inconsistency, as it does not have influence on the error once CCI_{fpt} is included.

The extent to which the models including similarity and its interactions are better suited to yield a consistent estimator for Γ than those excluding them depends on the extent to which similarity is correlated with other factors related to product addition and uncorrelated to firm-input-specific productivities, such as general technological changes that determine input-output coefficients. As there is not certainty about this extent, I decided to report results both with and without similarity and its interactions in all cases.

	D(future production)	D(future production)	D(future production)	D(future production)
D(quartile 2 of non-production workers)		-0.0069		-0.007
		(0.0354)		(0.0143)
D(quartile 3 of non-production workers)		-0.0006		-0.0006
		(0.0011)		(0.0011)
D(quartile 4 of non-production workers)		0.0019		0.0022
		(0.0041)		(0.0701)
Similarity		0.0170***		0.0111
		(0.0090)		(0.2133)
Similarity*D(quartile 2 of non-production workers)		0.0100*		0.0175**
		(0.0701)		(0.0116)
Similarity*D(quartile 3 of non-production workers)		0.0060		0.0258**
		(0.3741)		(0.0133)
Similarity*D(quartile 4 of non-production workers)		0.0179**		0.0592***
		(0.0358)		(0.0002)
log(CCI_{fpt})	-0.0061***	-0.0054***		
	(0.0000)	(0.0002)		
log($C\hat{CI}_{fpt}$)			-0.0193	-0.0217
			(0.4149)	(0.3524)
Constant	-0.0143	-0.0213	-0.1499	-0.2030
	(0.3308)	(0.2185)	(0.4047)	(0.2498)
Observations	22,508	18,809	9,738	9,738
R-squared	0.1857	0.2233	0.2630	0.2631

*** p<0.01, ** p<0.05, * p<0.1

Table 5. Results of regressions of product addition in $t+1$ as a function of a cost index and similarity in the use of materials interacted with dummies for quartiles of non-production workers. Estimated standard errors one line below the estimators.

The estimator for CCI_{fpt} is negative as expected under all the three specifications, both with and without the inclusion of the similarity and its interactions with the dummies for the quartiles of skilled labour. However, this estimator is statistically significant only when I use OLS without using the principal components of tariffs as instruments for CCI_{fpt} (it is actually highly significant in this case). The number below each estimator is the p-value for the null hypothesis that the corresponding true parameter is equal to zero. The p-values show that the null hypothesis of non-significance is far from being rejected when I use tariffs as instrument for CCI_{fpt}.

If the changes in tariffs are more exogenous than the prices of inputs to the non-modeled firm-product-specific factors that are correlated with product addition (as I reasonably assume), these results suggest that there is not conclusive causal evidence in favor of the model presented in the second chapter of this thesis. However, it is still valid to state that if changes in the prices of inputs were exogenous to such factors to some extent (as might be the case for firms in more competitive markets), the first two columns of Table 5 would indicate the presence of such evidence to some extent.

Very importantly, the results in Table 5 are highly consistent with the prediction derived from the proposition 2 in section 2 that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$. More specifically, the results in Table 5 are more consistent with this prediction than those shown in Table 2 in section 4. The only difference is that in Table 5 I included the CES cost index (CCI) as a regressor. As long as this variable is correlated with the product addition and its drivers (structural firm-input-specific productivities and input-output coefficients) are not fully captured by the similarity, its exclusion would generate an omitted variable bias in the results of Table 2 that would not exist in the results of Table 5 (provided that CCI is correlated with the similarity, as is the case). In summary, the results shown in Table 5 validate empirically the proposition 2 in section 2 to a larger extent than the results shown in Table 2, given the correction of a possible omitted variable bias in the former case.

The possibility of using $\hat{CCI}_{fpt}$ to get a consistent estimator of Γ in expression (5) depends on the validity of the tariffs as instruments. Formally, they must be uncorrelated with the errors $\epsilon_{fp,t+1}$. Given that expression (5) includes fixed effects by firm, year and product, $\epsilon_{fp,t+1}$ includes only factors that vary for two or more combinations of those three dimensions. In other words, it includes in theory all the firm-product-specific, firm-year-specific, product-year-specific and firm-product-year-specific factors that have effect on the product addition decisions and that are not included as regressors in expression (5). It is reasonable to assume that all the factors that vary to some extent across firms are uncorrelated with tariffs, as the generalized reduction in tariffs that took place in 2011 was a comprehensive policy that did not attempt primarily to attend requests of specific firms nor to solve firm-specific problems that might affect product addition.

Product-year-specific factors can be more problematic, as there may be time-varying product-specific factors that are common across all the firms that might potentially produce the product in question with two problematic properties:

(i) these factors might affect the firms' decisions to add the product in question, and (ii) these factors might be correlated with the tariffs imposed by the Colombian government to the imports from the U.S. of the inputs needed to produce that product. If (i) and (ii) were true, the tariffs would not be valid instruments.

The possible existence of time-varying product-specific factors with the properties (i) and (ii) above can be rationalized in several ways. Firstly, there might be time-varying product-specific shocks that might be correlated both with the product addition of the product in question and with the tariffs imposed by Colombia to the imports from the U.S. of the inputs needed to produce the product in question, such as demand shocks. If this were the case, the tariffs would not be valid instruments anymore. Secondly, some sectors [21] that produce some specific products might be able to foresee the future changes in the tariffs of the inputs needed for their production, and they might take this into account when deciding to produce their products or not some years before the tariffs reduction. If there is persistence in this "anticipation effect", the tariffs would not be valid instruments. Thirdly, the tariffs reduction might have been more pronounced for those products that are used as inputs of some specific products that are produced by sectors (this is, sets of firms) with specific properties, such as high lobbying capacity. If this capacity is persistent and correlated with product addition, the inputs tariffs would not be valid instruments.

The first possibility is the least worrisome, as I found in the first chapter of this thesis that the change in the average sales price of a product is uncorrelated with the probability that the product in question is added by the firms that may potentially produce it. This should not be the case if demand factors potentially correlated with the tariffs (such as demand shocks) affected the product addition. As for the second and third concerns, I do three things to discard their possible occurrence.

To start, I perform a parallel trends test for the product addition, in order to discard the possibility that the sectors that were more favoured by the inputs tariffs reductions added more products before such reductions than the rest (possibly because they foresaw these reductions). For this test, I use as pre-treatment period the years 2008 and 2009, and as post-treatment period the year 2010. This because the tariffs reduction was carried out in 2011. As for the treatment, I categorize as treated those products for which the weighted average of the tariffs on their inputs (using expenditures shares as weights) are above the median of this metric of inputs tariffs reduction, and as untreated the rest of products (this is, those for which this metric is below its median). The test consists of an augmented difference-in-differences regression. Formally, I ran the following regression:

[21] I use the term "sector" here to make reference to the set of firms that produce a product. I use this term when I need to emphasize the fact that these firms constitute an aggregate with possible specific characteristics such as lobbying capacity. I can use this term and the term "product" indistinctly without incurring into any conceptual mistake, as they both correspond to the same level of aggregation. Please notice that a firm can belong to several sectors here, as firms can be multiproduct.

$$D_{fp,t+1} = \alpha_0 + \alpha_1^{'} X_{fpt} + \alpha_2 D_{pt} + \alpha_3 W_f d_{t0} t + \alpha_4 W_f d_{t1} t + \gamma_p + \gamma_t + \epsilon_{fp,t+1} \quad (9)$$

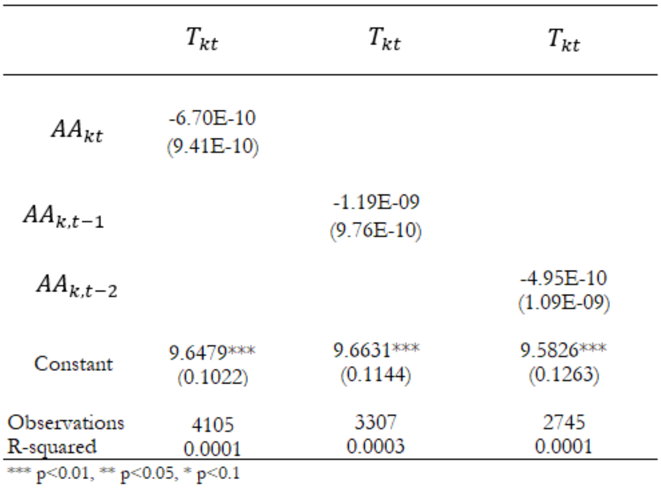
where $D_{fp,t+1}$ is the usual dummy for product addition of product p by firm f one year ahead, X_{fpt} is a set of covariates (the similarity index and its interactions with the quartiles of non-production workers), γ_p and γ_t are product fixed effects and time fixed effects, respectively, D_{pt} is the usual dummy in DID regressions that equals one if the observation corresponds to a treated product (this is, to a product that exhibits a metric of inputs tariffs reduction above its median) and the observation corresponds to 2010 and zero otherwise, w_p equals one if the observation corresponds to a treated product (this is, to a product that exhibits a metric of inputs tariffs reduction above its median), t is a trend, d_{t0} is one for 2008 and 2009 and zero otherwise, and d_{t1} is one for 2010 and zero otherwise.

The term $W_p d_{t0} t$ captures the difference in the linear trend before 2010 (namely, in 2008 and 2009) between the products more favoured by the reduction in the inputs tariffs and the products less favoured by this reduction. Therefore, if α_3 is statistically different from zero, there would be evidence that the trends were different, and it might be the case that sectors that foresaw that would be more favoured by the reduction of the tariffs of their inputs in 2011 started to produce their respective products to a larger extent than the rest of sectors several years before the tariffs reduction. The test consists then in performing a simple Wald test with the null hypothesis that $\alpha_3 = 0$ (parallel trends). I find a p-value of 0.5680 for this test, which allows me conclude that there is evidence of parallel trends in product addition of the sectors more favoured by the reduction of inputs tariffs and those less favoured by such reduction. This solves to some extent the second concern stated above (i.e., the concern of endogeneity because of a possible "anticipation effect").

In order to addess the third concern (possible endogeneity of inputs tariffs reductions because of phenomena such as lobbying capacity that may affect both product addition and tariffs reductions), I adapt the procedure used by Baccini et al. (2019) to the needs of this paper. Namely, I adapt it to take into account that in this paper it is the tariffs on inputs what matter, as these latter are the ones whose principal components I use as instruments. Namely, I run the following regression:

$$T_{kt} = \beta_{0,s} + \beta_{1,s} AA_{k,t-s} + \epsilon_{kt} \quad (10)$$

where T_{kt} is the tariff imposed by the Colombian government on the imports of input k from the U.S. in year t and $AA_{k,t-s}$ is the addition associated with input k in year $t-s$. It is calculated as follows:



	T_{kt}	T_{kt}	T_{kt}
AA_{kt}	-6.70E-10 (9.41E-10)		
$AA_{k,t-1}$		-1.19E-09 (9.76E-10)	
$AA_{k,t-2}$			-4.95E-10 (1.09E-09)
Constant	9.6479*** (0.1022)	9.6631*** (0.1144)	9.5826*** (0.1263)
Observations	4105	3307	2745
R-squared	0.0001	0.0003	0.0001

*** p<0.01, ** p<0.05, * p<0.1

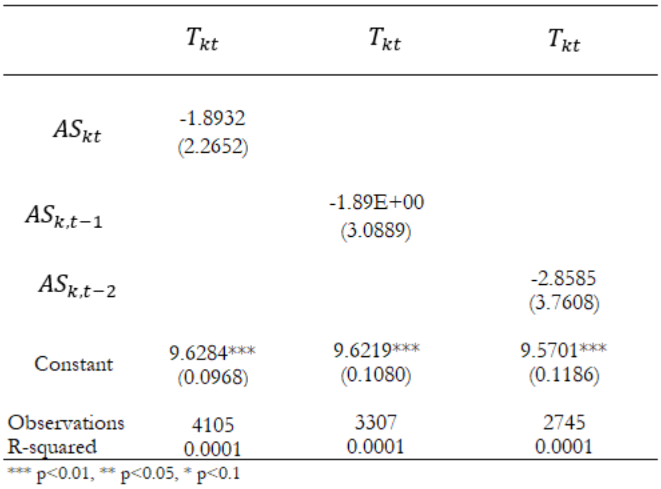
Table 6. Results of regressions of tariffs on the associated addition of products at the input level. Robust standard errors in parentheses.

$$\mathrm{AA}_{k,t-s} = \frac{\sum_{p=1}^{N_k} \Omega_{pk} \sum_f \mathbb{1}_{fp,t-s}}{\sum_{p=1}^{N_k} \Omega_{pk} N_p} (11)$$

where Ω_{pk} is the share of total expenditure of inputs to produce p that is spent on input k, $\mathbb{1}_{fp,t-s}$ is an indicator function that equals one if the potential new product p is added by firm f in $t-s$, N_k is the total number of potential new products that require input k for their production and N_p is the number of firms that have product p in their set of potential new products. Intuitively, $AA_{k,t-s}$ is a number between zero and one that measures the extent to which the products that require input k for their production are added by firms, weighting each of those products by the intensity in which k is used for its production. If I find β_1 to be negative and statistically different from zero, there would be evidence that the inputs that were intensively used in the production of the products chosen by the firms in the past exhibited larger tariffs reductions later, which might be indicative of lobbying capacity or similar phenomena (as long as such capacity manifests in higher product addition to some extent). The table 6 shows the results of the regression above for $s = 0, 1, 2$, in order to explore the relationship between tariffs and product addition for different spans.

I cannot conclude that β_1 is statistically different from zero in any case. This implies that inputs tariffs reductions were not more pronounced for inputs that yielded higher anticipated gains in terms of product addition, which supports the conclusion that non-observables such as lobbying capacity did not seem to determine the changes in tariffs. I run identical regressions with the additional inclusion of an autorregresive component of T_{kt}. The R2 increased to above 0.7. Very importantly, in this case I still cannot conclude that β_1 is statistically different from zero in any case.

Using the product addition as a possible predictor of tariffs might mask the effect of lobbying capacity or similar factors to some extent, as product addition



	T_{kt}	T_{kt}	T_{kt}
AS_{kt}	-1.8932 (2.2652)		
$AS_{k,t-1}$		-1.89E+00 (3.0889)	
$AS_{k,t-2}$			-2.8585 (3.7608)
Constant	9.6284*** (0.0968)	9.6219*** (0.1080)	9.5701*** (0.1186)
Observations	4105	3307	2745
R-squared	0.0001	0.0001	0.0001

*** p<0.01, ** p<0.05, * p<0.1

Table 7. Results of regressions of tariffs on the associated sales at the input level. Robust standard errors in parentheses.

may be influenced by several other factors. In order to surpass this problem to some extent, I repeated the regressions in Table 6 but using the sales instead of the dummies of product addition. Sales have two key properties: (i) I found this variable to be positively correlated with product addition in the first chapter of this thesis (which makes it an observable suspect to cause endogeneity, as I excluded it from the regression in (5)), and (ii) it is reasonable to assume that sectors with higher sales have higher lobbying capacity. With this in mind, I constructed the following variable of associated sales:

$$\mathrm{AS}_{k,t-s} = \textstyle\sum_{p=1}^{N_k} \Omega_{pk} Sales_{p,t-s}(12)$$

where $Sales_{p,t-s}$ is the logarithm of the total sales of product p in year $t-s$. Intuitively, AS measures how large were the sales of the products that require input k for their production, weighting each product by the extent to which k is required to produce it (this is, giving more importance to the products that are more intensive in k). Just as for the case of AA above, I ran regressions of T_{kt} as a function of $AS_{k,t-s}$ for $s = 0, 1, 2$. The results are shown in Table 7. I found the coefficient for AS to be statistically not different from zero in all cases. In words, the sectors with anticipated higher sales did not benefit from larger reductions of the tariffs of their most important inputs. In this case I also repeated the regressions with an autorregresive component of the tariffs. Once again, this did not alter the conclusion of statistical insignificance of the relevant parameter. This finishes my tasks to discard the third concern stated above.

Summarizing, it is reasonable to assume that the principal components of tariffs are valid as instruments for CCI_{fpt}, as it is reasonable to assume that they are uncorrelated with $\epsilon_{fp,t+1}$ in (5).

6 Evidence of the process of learning by using and of the effect of productivities on product addition

In this section I present an additional empirical analysis that explores the possible existence of empirical evidence in favor of the main assumption of the model proposed in the second chapter of this thesis, and also of its main mechanism. To do this I use the structural parameters (input-output coefficients and firm-input-specific productivities) that I identified and estimated in the previous section. The main conclusion from this section will be that the evidence is mostly in favor of the main assumption of the model presented in the second chapter of this thesis, and also of its main mechanism.

Main assumption and main mechanism of the model

The main assumption of the model proposed in the second chapter of this thesis is that there exists a process of learning by using, in which firms increase their firm-input-specific productivities as they use the respective inputs to a larger extent. The main mechanism of this model is that the products that require intensively for their production in t those inputs for whose use a firm f has high firm-input-specific productivities are more likely to be added by f in the years after t. Testing empirically the possible validity of this assumption and mechanism is very important, as I explain below.

What makes special (a) the process of learning by using (main assumption) and (b) the fact that firms add more easily products for which $\sum_k \phi_{pk}\phi_{fkt}$ is larger (main mechanism) is that if (a) and (b) happen in reality, then *necessarily* the finding that the correlation between similarity in t and the product addition after t persists over time is attributable to persistent firm-input-specific productivities to some extent, which is the main feature of the model proposed in the second section of this chapter. This would constitute additional and strong evidence in favor of the model presented in the second chapter of this thesis, in addition to the findings presented in the previous two sections, which are consistent with some of this model's propositions.

Let us assume for explanatory purposes that (a) and (b) occur in reality. This would necessarily imply that the main findings of this work are attributable to persistent firm-input-specific productivities because the interaction of (a) and (b) *necessarily* implies that the firms add more easily more similar products to their product mixes. If a firm f has a high productivity to use k in t, it optimally uses k intensively in t. If (a) happens in reality, then *necessarily* f will have a high productivity to use k in $t + 1$. If (b) happens in reality, then f *necessarily* adds to it product mix in $t + 1$ more easily those products that require k intensively for their production, and (once again) f uses k intensively to produce them, as it is optimal to do it. Therefore, the products produced by

f in t and the products produced by it in $t+1$ are *necessarily* similar in their input mixes, as they are both intensive in k. Therefore, my empirical finding that the products added by firms in the future have indeed input mixes that are more similar to the input mix used by firms in t is *necessarily* attributable to some extent to the persistent firm-input-specific productivities.

Econometric models

Fortunately, it is possible to run regressions to assess the empirical validity of (a) and (b). As for (b), it is possible to run product-specific regressions of dummies of product addition as a function of $\sum_k \phi_{pk}\phi_{fkt}$ in order to assess the empirical validity of (b). If the estimator for the coefficient for $\sum_k \phi_{pk}\phi_{fkt}$ were positive, I could claim to have found evidence consistent with (b), as this would mean that firms have on average a higher probability of adding products that are intensive in the inputs for whose use they have high productivities. Formally, I run the following regression separately for each p:

$$D_{fp,t+1} = \delta_{0,p} + \delta_{1,p} \sum_k \phi_{pk}\phi_{fkt} + E_{fp,t+1} \tag{13}$$

$D_{fp,t+1}$ equals one if p is added by f in $t+1$, and zero otherwise. This variable is defined on the domain of all the potential new products for f [22]. Parameters ϕ_{pk} and ϕ_{fkt} were estimated in section 5. $E_{fp,t+1}$ is an error term that is assumed to have mean zero and to follow a pattern of variability and correlation across firms and products that is directly estimated from data [23].

If δ_1 is found to be positive for a product p, this would mean that p is added more easily on average by firms with high productivities for the use of the inputs in which the production of p is intensive. If this were the case, this evidence would not be contradictory with (b) above. If this were the case for many products, the evidence would be mostly in favor of (b). I ran a different estimation for each product. This is why both coefficients in (13) are product-specific.

As for (a), I ran input-specific regressions of the firm-input-specific productivity of each firm f in material k in year $t+1$ ($\phi_{fk,t+1}$) as a function of the amount of k used by f in year t ($_{fkt}$). Formally, I ran the following regression:

$$\phi_{fk,t+1} = \theta_{0,k} + \theta_{1,k} M_{fkt} + E_{fk,t+1} \tag{14}$$

$E_{fk,t+1}$ is a firm-material-year-specific error term that is assumed to have the same properties as $E_{fp,t+1}$ above. A positive estimator of θ_1 would mean that it does happen on average for input k that the productivities of firms to

[22] See section 2 for a definition of the potential new products

[23] In short, this means that I use the Huber-White sandwich estimator of the variance-covariance matrix

use it grow on average as they use it more (by the process of learning by using). This is exactly what (a) states. If this were the case for many inputs, I would have found evidence that is generally consistent with (a). I ran a different estimation for each input. This is why both coefficients in (14) are input-specific.

Distribution of the productivities and of the input-output coefficients

There are two key variables involved in the expressions (13) and (14) that are not directly observable from data: ϕ_{pk} and ϕ_{fkt}. I found estimators for them in section (5). Figures 7 and 8 show the distributions of the estimators of ϕ_{pk} and ϕ_{fkt}, respectively.

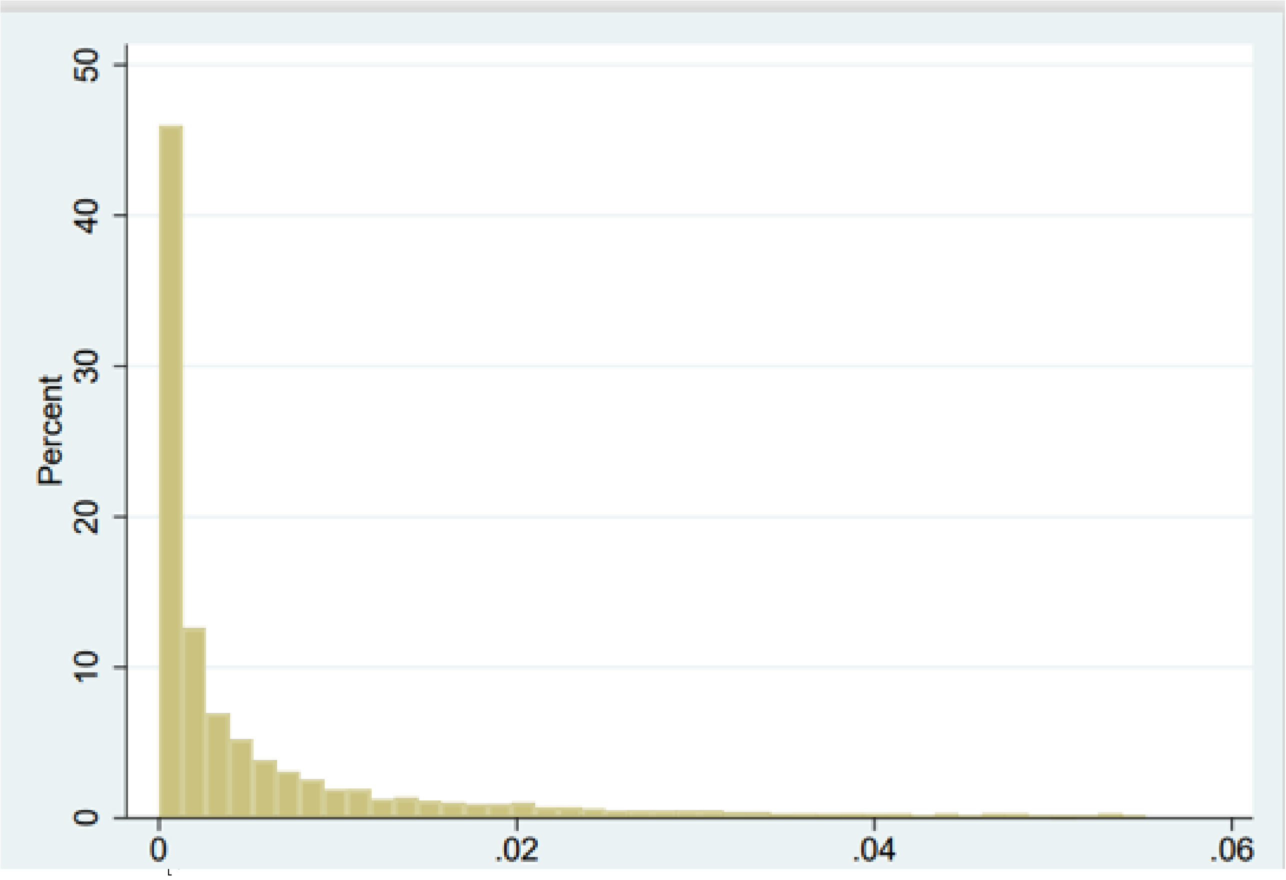


Figure 3. Histogram of estimators of input-output coefficients ϕ_{pk}. It shows the percentage of each interval.

The estimators of ϕ_{pk} and ϕ_{fkt} shown in Figures 3 and 4 are both concentrated in low values. The distribution is in both cases highly skewed towards the lowest intervals. There are also some firms and products with intermediate values of firm-input-specific productivities and input-output coefficients, respectively. Finally, there are a few firms and products with high firm-input-specific

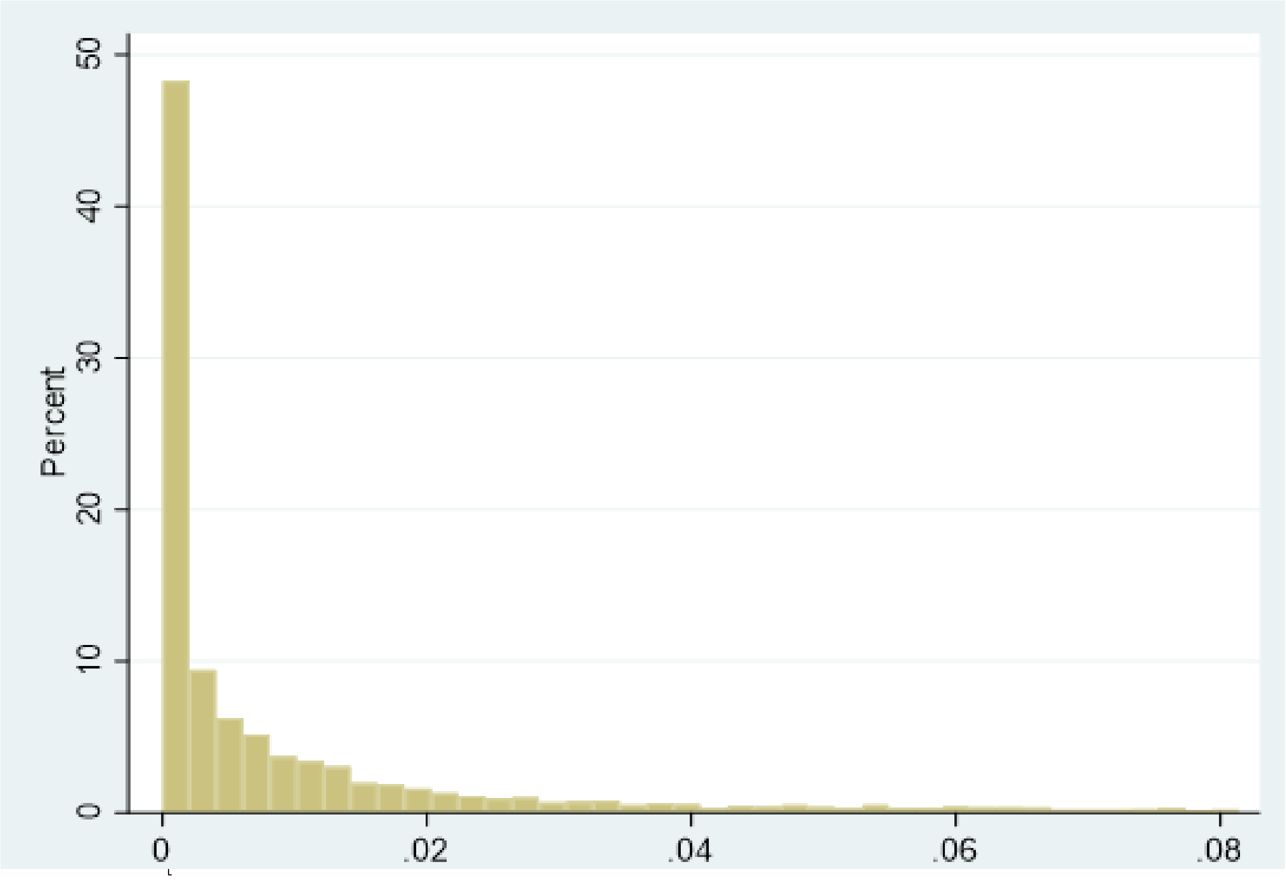


Figure 4. Histogram of firm-input-specific productivities ϕ_{fkt}. It shows the percentage of each interval.

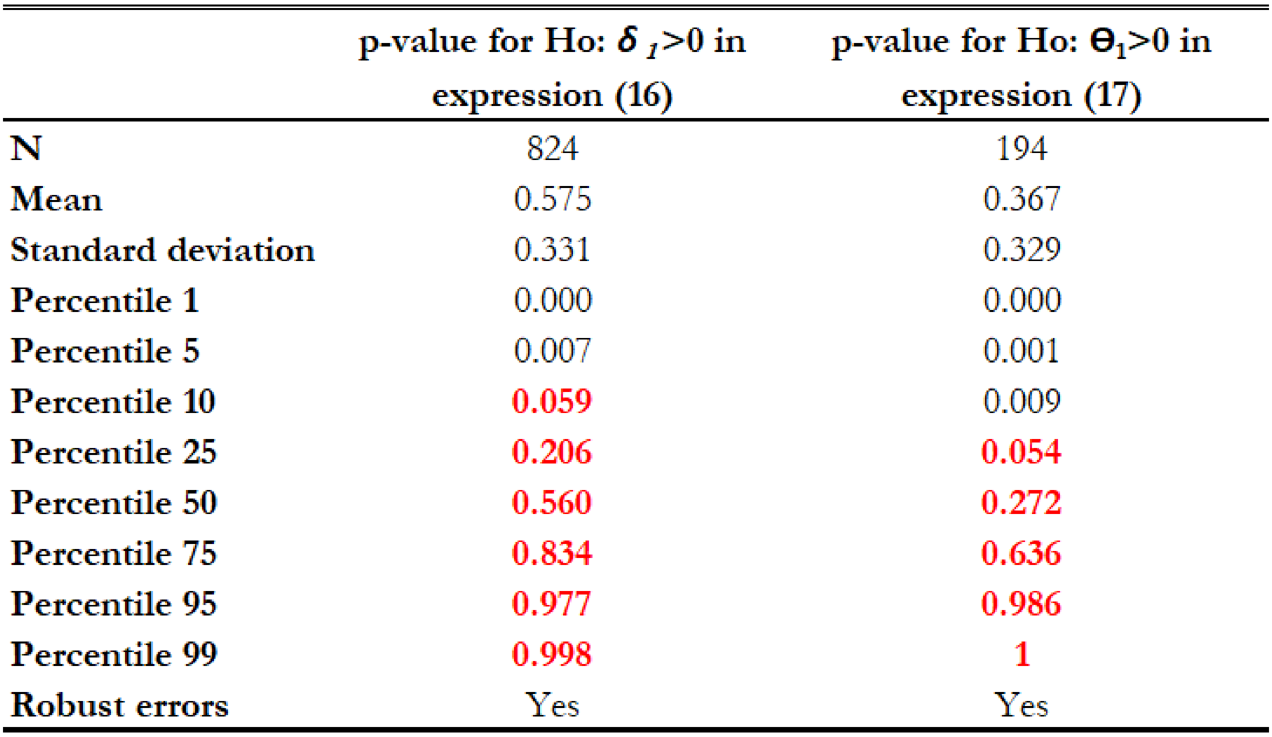
productivities and input-output coefficients. In summary, the contributions of inputs to production and the firm-input-specific productivities of firms are both small in most cases, and there are just a few cases in which inputs contribute to a large extent to production or firms are very productive at using inputs. This is qualitatively consistent with the finding from Del Gatto et al. (2006), who found that productivity has an empirical distribution that can be characterized as a theoretical Pareto distribution, as this latter is also skewed towards low values, and exhibits very high values at its upper tail.

Results

I ran all the feasible product-specific and input-specific regressions of expressions (13) and (14), respectively. By feasible I mean that I ran all those regressions for which I had a sufficient number of observations. This restriction allowed me to run 824 product-specific regressions and 194 input-specific regressions, as shown in Figure 5. As the estimators can be not compared across products and inputs, they are not shown here. Instead, I present in Figure

5 summary statistics that characterize in each case the distribution of the p-values for the null hypothesis that the respective true parameter is positive. More specifically, left column of Figure 5 shows the percentiles and other statistics of the 824 p-values of the 824 test statistics for the null hypothesis that δ_1, p is positive for the 824 products for which I ran regressions for expression (13). Analogously, right column of Figure 6 shows the percentiles and other statistics of the 194 p-values of the 194 test statistics for the hypothesis that θ_1, k is positive for the 194 products for which I ran regressions for expression (14).

Positive values of true parameters δ_1, p and θ_1, k would constitute evidence in favor of (a) and (b). The main conclusion from Figure 5 is that this is mostly the case. More specifically, left column shows that for more than 90 percent of products the evidence does not allow a rejection of the hypothesis that the probability of addition is higher for the products that are intensive in the inputs for whose use firms have higher productivities ((b) above). Similarly, right column indicates that for more than 75 percent of inputs the evidence does not allow a rejection of the hypothesis that the firm-input-specific productivities grow as the respective inputs are used to a larger extent. In summary, the evidence is mostly in favor of (a) and (b).



	p-value for Ho: δ_1>0 in expression (16)	**p-value for Ho: Θ_1>0 in expression (17)**
N	824	194
Mean	0.575	0.367
Standard deviation	0.331	0.329
Percentile 1	0.000	0.000
Percentile 5	0.007	0.001
Percentile 10	**0.059**	0.009
Percentile 25	**0.206**	**0.054**
Percentile 50	**0.560**	**0.272**
Percentile 75	**0.834**	**0.636**
Percentile 95	**0.977**	**0.986**
Percentile 99	**0.998**	**1**
Robust errors	Yes	Yes

Figure 5. p-values of product-specific and input-specific estimators of slopes of expressions (9) and (10)

7 Conclusions

This paper uses a random sample of a comprehensive dataset of Colombian manufacturing firms to analyze empirically the phenomenon of product addition and its determinants at the firm level. Its main conclusion is that the analyzed empirical evidence is mostly in favor of the most important predictions of the theoretical model of product addition and persistent firm-input-specific productivities proposed in the second chapter of this thesis.

To start, the main regressions of this paper yield evidence in favor of the first proposition of the model presented in the second chapter of this thesis. Namely, I found that firms add on average more easily in the future new products that are more similar to their current production in terms of the input mix, and that this correlation is stronger in the firms with the highest levels of skilled labour (proxied by the number of non-production workers). Very importantly, I found stronger empirical support for the effect of skilled labour on the correlation between similarity and product addition after correcting a possible omitted variable bias caused by the exclusion of the potential cost. I also found evidence that the correlation between current similarity and future product addition remains over time. In other words, the current similarity in the use of inputs is positively correlated with product addition in the future, even 5 years ahead for the firms with the highest values of non-production workers.

I interpret these results as primary evidence in favor of the key role of the interaction of firm-input-specific productivities and input-output coefficients as in the model I proposed in the second chapter of this thesis. More specifically, in that model, firms accumulate over time persistent productivities to use specific inputs as they use these inputs in their productive processes, and this allows them to add more easily in every year products that use those inputs intensively.

In order to establish if the findings explained so far can indeed be attributed to the persistence of firm-input-specific productivities, I do two additional things in this work. Firstly, I used a source of exogenous variation in the prices of inputs to analyze possible causal evidence in favor of the model proposed in the second chapter of this thesis, which relies on these persistent firm-input-specific productivities. Secondly, I used structurally estimated firm-input-specific productivities and input-output coefficients to test empirically the validity of the most important assumption of the model proposed in the second chapter of this thesis, and also of its main mechanism.

For the analysis of causality I use a generalized unilateral reduction of tariffs to the imports from the U.S. carried out by the Colombian government in 2011. This reduction was deepened in the subsequent years. The model proposed in the second chapter of this thesis predicts that this general reduction of the prices of inputs should have affected the product addition of different potential new products by different firms to different extents, given the pre-change firm-input-specific productivities and the input-output coefficients for each firm-product combination. The granularity of this prediction allows me to test for possible causal evidence in favor of the model proposed in the second chapter of this thesis by running a regression for a firm-product-year dummy for product addition that incorporates a structural variable that accounts for the firm-product-specific effect mentioned above. The estimator for this variable is negative but it is not statistically significant in the specifications that correct for possible endogeneity. I conclude from this that there does not exist conclusive causal evidence in favor of the model proposed in the second chapter of this thesis.

As for the direct testing of the main mechanism and the main assumption of the model presented in the second chapter of this thesis, I use structurally estimated parameters to run regressions whose coefficients capture such mechanism and assumption. The results are mostly in favor of the existence of these latter.

In summary, I find in this paper primary evidence in favor of the explanation provided by the second chapter of this thesis for the phenomenon of product addition by firms. Such explanation states that firms add more easily products that are intensive in the inputs for whose use they have higher firm-input-specific productivities. As these productivities remain over time to some extent, they lead firms to produce products that are similar across years in terms of their input mixes. In order to establish additional and more solid evidence in favor of this explanation, I carry out a causality analysis and a structural estimation of key parameters of the model, which I subsequently use also to test directly the empirical validity of the key mechanism behind such explanation, and of the main mechanism of the model. This latter task yields evidence in favor of the model presented in the second chapter of this thesis, whereas the causality analysis does not yield conclusive results.

The results summarized here can be used to inform the process of design and implementation of growth policies. This because these policies should aim at boosting the process of product addition at the firm level, as this would result in gains in terms of efficiency and growth. If the agencies in charge of promoting growth decided to effectively boost the process of product addition at the firm level, they would face the key and non-trivial question of how to do this. The results found in this paper contribute to answering this question to some extent.

The main findings from this paper can be summarized as follows in terms of the dimensions that matter for public policies and that may be affected by these latter: (a) There is path dependence in the process of product addition by firms, as they move more easily to more similar products in terms of the input mix. (b) the capacities of firms to use different inputs determine both the paths chosen by them to expand their product mixes and the speed at which they go through these paths. (c) Having more skilled labour increases the speed at which a firm goes through its path, whatever this latter is. (d) Paths are reinforcing: if a firm chooses a path of products that are intensive in an input, this decision makes this firm even more prone to keep going through this path. (e) Given (d), reaching paths that contain less similar products may take firms longer, unless their capacities to use the inputs in which these products are intensive increase because of a reason that is external to the firm (such as a public policy).

Given this context, there are two main courses of action that national and sub-national authorities in charge of designing and implementing growth policies can take. Firstly, they can implement policies that increase the speed at which firms go through the paths they choose. Secondly, they can help firms to reach new paths of diversification (through product addition).

In order to increase the speed at which firms go through the paths they optimally choose, the authorities can do three things. Firstly, they can facilitate the hiring of the relevant skilled labour by firms. I found in this thesis that having more skilled labour in general increases the speed at which firms go through their optimal paths, but it remains a pending task for each authority to establish what types of skilled labour are critical in each context. Once this is done, the authorities can implement programs that subsidize the hiring of these critical skilled workers conditional on them working on activities that are closely related to product addition. Secondly, the authorities can implement policies that help firms to maintain their existing capacities to use inputs to larger extents. This can be done by improving learning protocols and standardization, facilitating the spreading of input-specific knowledge within and across firms and mitigating the losses of knowledge caused by the turnover of production workers. Thirdly, the authorities can focus their training programmes on identifying the input-specific capacities of each firm, and then on increasing these capacities. Product innovation programmes can focus on activities that aim at implementing new ways of using the inputs identified as critical for each firm.

On the other hand, authorities might be interested in helping firms to reach new paths of diversification. This might be the case if the authorities identify key new industries that they want the country to develop for different economic or non-economic reasons, but that cannot be developed in the desired time horizon with the current state of capacities of firms to use their inputs. In short, authorities might be interested in helping firms to transit from their current paths to other "better" paths (whatever "better" means in each context) in an established time horizon. This thesis suggests possible courses of action to get this. Authorities can implement policies to increase exogenously the input-specific capacities of firms to use inputs that they do not use as frequently and proficiently as others. This can be done by prioritizing the access of firms to these inputs and their productive experimentation with them in the training programmes and also in the knowledge and technology transfer programmes. Partial subsidies to the use of the inputs identified as critical for the key new industries can be also implemented. This policy should yield tangible results, as I found that the costs of production are indeed correlated with product addition, as expected. This thesis does not imply that policies that aim directly at promoting the development of new products (without focusing on the use of the inputs that are critical for their production) are incorrect, but only that focusing on increasing the firms' capacities to use critical inputs can effectively help firms to produce those new products.

Path dependence and the intuition that skills determine what paths are chosen by countries and firms and how easily these go through their chosen paths are not exclusive features of this work. Hausman and Hidalgo (2008), Feenstra and Rose (2000) and several other authors arrived to very similar conclusions and conjectures (please see the first chapter of this thesis for a more detailed discussion about this topic). However, this work offers more specific findings by analyzing more granular phenomena, and this allows me to yield more specific recommendations. Namely, my findings about the existence of persistent firm-input-specific productivities, their dependence on the actual use of the respective inputs, the effect of skilled labour on their persistence and their effect on product addition allows me suggest more specific courses of action, as I did in the previous paragraphs. Future work might focus on analyzing more precisely what types of skilled labour are critical in each context and what policies allow firms to maintain and boost their firm-input-specific productivities to larger extents in each context.

References

Angrist, Joshua D., and Pischke, J. "Mostly Harmless Econometrics: An Empiricist's Companion", *Princeton: Princeton University Press.*

Baccini, Leonardo, Impullitti, Giammario and Malesky, Edmund J. (2019). "Globalization and state capitalism: Assessing Vietnam's accession to the WTO", *Journal of International Economics*, 119: 75-92.

Boehm, Johannes, Dhingra, S. and Morrow, J. (2019). "The Comparative Advantage of Firms", *CEPR discussion papers*, 13699.

Cameron, Colin, Gelbach, J. and Miller, D. (2008). "Bootstrap-Based Improvements for Inference with Clustered Error", *The Review of Economics and Statistics*, 90 (3): 414-427.

Correia, Sergio (2016). "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator", *Working Paper.*

Del Gatto, Massimo, Mion, G. and Ottaviano, G. (2006), "Trade Integration, Firm Selection and the Costs of Non-Europe", *Mimeo, University of Bologna.*

Ding, Xiang. (2020), "Industry Linkages from Joint Production", *Unpublished.*

García García, J., Montes, E. and Giraldo, I. (editors) (2019), "Comercio Exterior en Colombia. Política, Instituciones, Costos y Resultados", *CEP, Banco de la República de Colombia.*

Klette, Tor Jakob (1994), "R and D, Scope Economies, and Plant Performance", *The RAND Journal of Economics*, Vol. 27, No. 3, 502-522.

MacDonald, J. (1985), "R and D and the Directions of Diversification", *The Review of Economics and Statistics*, Vol. 67.

Melitz, Marc, and Ottaviano, G. (2008). "Market size, Trade, and Productivity", *Review of Economic Studies*, Vol. 75, 295-316.

Sutton, John (2012). "Competing in capabilities. The Globalization Process", *Oxford University Press.*

Product addition and comparative advantages in Colombian manufacturing firms in a context of trade liberalization

Andres Trejos

June 19, 2025

Abstract

This work presents new empirical evidence of possible firm-level and firm-product-level determinants of the important phenomenon of product addition at the firm level. I do this by using a stratified sample of a rich panel of Colombian manufacturing firms from 1992 to 2017. Importantly, this panel includes all the firms with 10 or more employees and/or sales above 120k USD per year. I find that firms tend to add to their product mixes products with similar input requirements (input mixes hereinafter) as products already in their portfolios. In addition, I find in general that this correlation between product addition and similarity in the input mix is stronger on average in firms with more skilled labour (proxied here by the number of non-production workers), especially after controlling for a possible omitted variable bias. These findings are consistent with the theoretical hypotheses presented in the second chapter of this thesis. In addition, I find that the correlation between future product addition and the current similarity in the input mix remains positive and significant even 5 years ahead, but I do not find evidence that less similar products are added each year as I move further into the future, as I hypothesized in the second chapter of this thesis. In addition, I use product-specific exogenous unilateral reductions of tariffs to imports from the U.S. implemented in 2011 by the Colombian government to analyze possible causal evidence in favor of the model presented in the second chapter of this thesis, which I cannot establish unambiguously. Finally, I explore possible empirical evidence in favor of the most relevant mechanism of this model and of its main assumption. Its main mechanism is that those products that are intensive in the inputs in whose use a firm is proficient are more profitable for this firm. Its main assumption is that the firm-input-specific productivities of a firm (which measure its proficiency to use the different inputs) grow over time as this firm uses the respective inputs to a larger extent. The empirical evidence is mostly consistent this mechanism and assumption.

1 Introduction

Product addition at the firm level is a key phenomenon for economies to achieve economic efficiency and growth. As I explained in detail in the first chapter of this thesis, this phenomenon has been found to be the main driver of the process of product diversification at the firm level, which in turn is an important source of reallocation of resources within firms towards their more efficient use, as it allows firms to move their resources to more profitable activities (by switching products). Given this, product diversification contributes as much as the phenomenon of firm entry and exit to the evolution of U.S. aggregate manufacturing output. It is also an important determinant of variability in the aggregate economic activity [1]. In spite of this, the determinants of the process of product addition at the firm level have not been sufficiently analyzed. This paper contributes to closing this gap by presenting new facts about the possible determinants of this process for the case of Colombian manufacturing firms.

By using a very rich dataset of Colombian manufacturing firms, this paper does three things. Firstly, it explores the possible nonlinear relationship between product addition at the firm level and several firm-level and firm-product-level possible determinants. Namely, I explore the possible empirical validity of the first theoretical proposition from the second chapter of this thesis that those potential new products that require an input mix that is more similar to that used in a year by a firm are more easily added by this latter in subsequent years. Moreover, I also explore the possible empirical validity of the second theoretical proposition in the second chapter of this thesis that this relation between product addition and similarity is larger in firms with more skilled labour, and that firms with more skilled labour add more easily potential new products to their product mixes (all else constant).

Secondly, I explore the possible correlation between product-level costs of production and the probability of product addition. For this, I use as instruments the tariffs imposed by the Colombian government to the imports from the U.S. during a generalized and unilateral reduction of these tariffs. As this generalized reduction in tariffs is exogenous to the decisions of product addition of the Colombian manufacturing firms and it is precisely costs of production what cause the role of similarity as a determinant of product addition in the theoretical model of the second chapter of this thesis [2], this exploration constitutes an examination of possible causal evidence in favor of the theoretical model presented in the second chapter of this thesis. For this analysis I use the same dataset of Colombian manufacturing firms mentioned before.

Thirdly, I explore the possible empirical validity of the main mechanism of the theoretical model from the second chapter of this thesis, and also of its main assumption. Namely, I explore possible evidence that the interactions between input-output coefficients and firm-input-specific productivities determine the process of product addition by firms. In order to analyze the empirical validity of this mechanism, I explore the possible correlation between the probability of addition and the aforementioned interaction. For this analysis I use the same dataset of Colombian manufacturing firms mentioned before. As for the main assumption (which I named ”learning by using in the second chapter of this thesis), I explore the possible correlation between the use of inputs by firms in

[1] Please see the first chapter of this thesis for a detailed explanation of the importance of product diversification and for the references on which this explanation is based.

[2] Please see the third theoretical proposition in the second chapter of this thesis for details.

a given year and the corresponding firm-input-specific productivities one year ahead.

As for the first task mentioned above, I find a positive and statistically significant correlation between similarity in the use of inputs and product addition one year ahead. I also find that such correlation is stronger in the firms with the highest quantities of non-production workers (which I use as a proxy of skilled labour). Very importantly, my results for the effect of non-production workers on the correlation between similarity and product addition improve when I include the potential cost of production as a regressor. This suggests a possible omitted variable bias in the specifications that exclude this variable. In addition, I find that the positive correlation between product addition and similarity remains significant for longer time spans, although its size does not decrease monotonically as I move further away in the future, as expected. In summary, this paper presents evidence that is mostly consistent with the first proposition in the second chapter of this thesis, and partially in favor of its second proposition (especially when the potential cost of production is included as a regressor).

As for the second task mentioned above (called here "causality analysis"), the results presented here do not conclusively establish or deny causal evidence of the validity of the theoretical model presented in the second chapter of this thesis. This because I explore several specifications with and without instrumented regressors, and the conclusions are heterogeneous (although I do find in all cases a negative point estimator for the cost of production as a regressor for the probability of addition as dependent variable, as expected). Finally, I find empirical evidence that is mostly consistent with the possible existence of the main mechanism of the model presented in the second chapter of this thesis, and also evidence that is mostly consistent with the validity of its main assumption.

It is worth to mention that I perform in this paper a structural estimation of the structural input-output coefficients and of the firm-input-specific productivities, which are then used to carry out the second and third tasks mentioned above. Namely, I use them for the causality analysis and for the empirical examination of the main mechanism and the main assumption of the theoretical model presented in the second chapter of this thesis. For this, I use the first-order conditions of the product-specific conditional profits of a firm with respect to the different inputs, which I derived in the second chapter of this thesis. A log-linearization of these conditions yields a linear estimable expression in which both the input-output coefficients and the firm-input-specific productivities are identified.

This paper is closely related to the very recent and far from conclusive branch of literature that analyzes empirically the possible determinants of the phenomenon of product addition at the firm level. The papers that belong to this branch share two key features with the work presented in this paper. Firstly, their authors test the empirical validity of specific theoretical predictions from theoretical models that are proposed by themselves, instead of performing agnostic regressions in the search for possible significant correlations. Secondly, their authors carry out a causality analysis or a complementary empirical analysis that attempts to establish empirical evidence in favor of their main mechanisms. Very importantly, the conclusions from these works are contradictory to some extent, as I explain below. This highlights the importance of the work presented here, as it brings new evidence into this nascent discussion and sheds light about possible mechanisms that might lie behind some results from the past.

Ding (2020) used U.S. data to test his prediction that it is similarity in the use of knowledge inputs (and not in the general use of inputs) what matters for product addition, and to test for the validity of the main mechanisms involved in his model. In contrast, Boehm et al. (2019) used Indian data to test for their prediction that the probability of product addition is higher for more similar products in terms of the *general* use of inputs (and not just of the use of *knowledge* inputs), and to test for the empirical validity of the main mechanisms involved in their model. They also used the exogenous process of dereservation in India to establish causal evidence in favor of their model.

My work is very similar to the one carried out by Boehm et al. (2019) both in terms of the nature of the datasets and of the direction of the results, as they also explored the possible existence of an empirical positive correlation between similarity in the input mix and product addition, and they also found evidence of this existence. However, the analyses presented here explore the role of additional characteristics of firms in the process of product addition. Namely, I explore both the empirical correlation between product addition and skilled labour, and also the way in which skilled labour determines the size of the correlation between similarity in the input mix and product addition. In short, I find that the amount of skilled labour used by a firm determines the size of the correlation found by Boehm et al. (2019) and by myself between product addition and similarity in the input mix. This role of skilled labour might be actually related to Ding's results, as it might be indicative of the importance of knowledge inputs (which are closely related to skilled labour). I do not explore this possibility in this paper.

Given the aforementioned importance of the process of product addition at the firm level for growth and efficiency, identifying its determinants is important for growth policies. This because growth policies should ideally aim at improving the state of firms in the dimensions that are identified as significant determinants of product addition, as this latter phenomenon is in turn a source of growth and efficiency. As this work identifies some of these determinants, its conclusions may be used to inform the process of design and implementation of growth policies. This is why I include a brief discussion of the main possible policy implications of the findings of this paper in the last section.

This document has seven sections. The first of them is this introduction. The second section presents a brief description of the theoretical model proposed in the second chapter of this thesis and states its three propositions, as I refer to this model and these propositions often throughout the rest of this document. The third section describes in detail the dataset to be used in the subsequent sections. The three subsequent sections present the results of the first, second and third tasks described above, respectively. Namely, the fourth section presents all the results of the regressions of product addition on similarity in the input mix, skilled labour and the interactions between these two latter. The fifth section presents the results of the causality analysis. The sixth section presents the results of the analysis of the empirical validity of the main mechanism and the main assumption of the theoretical model presented in the second chapter of this thesis. Finally, the seventh section presents the conclusions of this work and its main possible policy implications.

2 The theoretical model and its propositions

I propose in the second chapter of this thesis a static model of partial equilibrium with dynamic implications. In this model all firms are potentially multiproduct. Each firm maximizes its profits in a year t by choosing its product mix, its produced quantity of each produced product, the quantities of labor and capital to be used in its production process and the quantities of all the inputs to be used for the production of each produced product [3]. Very importantly, this use of inputs in t determines the firm-input-specific productivities in $t+1$. More specifically, there is a process of *learning by using* in which a firm f's firm-input-specific productivity to use a particular input k grows more between t and $t+1$ the more k is used by f in t. This implies that there is persistence in the firm-input-specific productivities in this model. These latter do not only not disappear from one year to other, but they may grow as a consequence of the inputs use.

Given this structure, this theoretical model yields naturally three key results. Firstly, a firm f adds more easily in the future those potential new products that require input mixes that are more similar to the input mix that f uses in the present. This result comes from the fact that f finds it more profitable to produce in t those products that require intensively those inputs for whose use f has a high productivity (let us call these group "f's preferred inputs"). In turn, the production of these effectively produced products requires intensively the use of f's preferred inputs, as they are by definition intensive in these latter. This intensive use of f's preferred inputs makes f even more productive at using f's preferred inputs in $t+1$ (by the process of learning by using), which implies that f will choose once again in $t+1$ potential new products that are intensive in f's preferred inputs to be added to its product mix (given their higher profitability for f).

As a consequence of the dynamics described above, there is a higher similarity in the intensity in different inputs (input mix) between the products produced in t and the new products produced in $t+1$ than between the former and the potential new products *not* produced in $t+1$. This because the new products produced in $t+1$ are intensive in f's preferred inputs (just as the products produced by f in t), whereas the potential new products *not* produced

[3] Here the term "inputs" does not make reference to capital or labour, but only to the other physical inputs used in the production process that are directly transformed into outputs. Please see the first chapter of this thesis for a detailed explanation of this differentiation and of its relevance.

in $t+1$ are not intensive in f's preferred inputs (that is precisely why they are not chosen to be produced). As f's productivities to use its preferred inputs keep on growing in time, this higher similarity for the produced potential new products than for the not produced ones persists in all the subsequent years. In short, firm-input-specific productivities determine the profitability to produce different products. As these productivities remain in time, the products produced in different years are similar to each other in terms of their input mixes (as they are all intensive in the same inputs).

This relationship between product addition and similarity changes over time in the model presented in the second chapter of this thesis. If a product p's input mix is not sufficiently similar in t to the input mix used by a firm f in t as to be added to f's product mix in $t+1$, it may be eventually added in a subsequent year, all else constant. This because firm-input-specific productivities may grow mechanically over time in this model (although they grow to different extents, depending on how intensively they are used). In other words, all the products may eventually be added (sooner or later) by all firms to their product mixes, all else constant. However, how soon a product p is added to a firm f's product mix depends positively on how similar is p's input mix to the input mix used by f in t. Namely, the less similar products are added later, as I explain below.

The less similar products may be added later by a firm f to its product mix because it may take longer to f to reach the firm-input-specific productivities that are needed to produce these products. This happens because of two reasons. Firstly, f is not proficient in t in the inputs that are intensively used to produce p (if it were, it would use these more intensively and its input mix would be more similar to p's one in t). This implies that f does not use those inputs intensively in t. This low intensity in the use of these inputs in t makes the increase from t to $t+1$ in f's productivity to use them less pronounced than the increase expected for other inputs. Secondly, this latter fact makes the firm once again less intensive in these inputs than in others in $t+1$, and so on. As a consequence, less and less similar products are added to the product mix of a firm every year.

All the pieces of intuition presented above are summarized in the first proposition of the second chapter of this thesis, which states the following[4]:

Proposition 1 (main proposition): *Firms are more likely to add in the future potential new products whose input mixes are closer to the firms' current input mixes. The further one moves away from the present time into the future, the less similar are the input mixes of the potential new products that are added to firms' product mixes.*

As I explained above, I assume that the process of learning by using is more pronounced in firms that use more skilled labour. Formally, the increase between t and $t+1$ in f's productivity to use an input k that occurs when k is

[4] Please see the second chapter of this thesis for a formal proof of this proposition and of the propositions 2 and 3 below, and for a formal definition of potential new products and of similarity in the input mix

more used in t is higher if f uses more skilled labour in t[5]. As a consequence of this, two results arise. Firstly, all the potential new products are more easily added in all the years after t by firms with more skilled labour. Intuitively, these firms exhibit larger increases in all the firm-input-specific productivities between all the pairs of contiguous years from t on. Therefore, all the products are more profitable for them in all the years.

Secondly (and more importantly), this advantage of the firms with more skilled labour favors to a larger extent the products that require input mixes that are more similar to the input mix used by those firms. In other words, the advantage of the more similar products stated in the proposition 1 above is more pronounced in firms with more skilled labour.

The important fact stated in the previous paragraph comes from the key feature of the model proposed in the second chapter of this thesis that skilled labour interacts *multiplicatively* with the use of each input in t in the equation that characterizes the process of learning by using, which determines the firm-input-specific productivities in $t + 1$. This multiplicative nature of this interaction implies that the skilled labour boosts the productivity of a firm f to use a particular input k in the future to a larger extent if f is used to a larger extent by f in t. As f's preferred inputs are used to a larger extent by f in t, these inputs are more favored (in terms of f's productivities to use them) by the skilled labour's effect on the process of learning by using. This differential effect in favor of these inputs increases to a larger extent the future profitability of the potential new products that require them intensively. As I explained above, these latter products are the ones effectively produced by the firm in the future, and they have a higher similarity with f in terms of the input mix.

These two results from the model presented in the second chapter of this thesis are summarized and materialized in the following proposition:

Proposition 2 (the role of skilled labour): *Firms with more skilled labour are more likely to add new products in the future. The advantage of the more similar products stated in Proposition 1 is greater in firms with more skilled labour.*

The theoretical model presented in the second chapter of this thesis also predicts two things about the effects of a generalized reduction in the prices of inputs in the present (year t). Firstly, there occurs a heterogeneous increase in the profitability of addition of the different products by the different firms. More specifically, this increase is higher for the firm-product combinations for which the corresponding products intensively require for their production those inputs whose prices fall more. Secondly, this compound differential effect is even larger if the corresponding firms have high firm-input-specific productivities in the inputs whose prices fall more.

The intuition for the second prediction above is as follows: assume that a

[5] Please see the second chapter of this thesis for an explanation of the soundness of this assumption.

firm f is better at using (say) glass than (say) paper and the opposite is true for another firm $f2$. Also assume that the price of glass falls more than that of paper in t. Given its proficiency to use glass, f is able to take more advantage of the pronounced reduction in the cost of glass in t than $f2$, and f uses glass to a larger extent in this year. As a consequence, f's productivity to use glass in subsequent years increases more than that of $f2$ from t to $t+1$ by the process of learning by using. In turn, this increases f's profitability to produce glass-intensive products in all the subsequent years to a larger extent than for other firms [6].

As for the intuition for the first prediction, it is as follows: if a product p is intensive in the inputs whose prices fall more, its cost of production falls more than that of the other products in t, and therefore the profitability of its production in t increases more than that of other products. Given this, more firms produce p in t than other products, all else constant. As a consequence, the inputs that are used intensively to produce p (this is, the same whose prices fell more) are demanded to a larger extent in t than other inputs, all else constant. This increases the productivities of firms to use those inputs in the future more than the productivities to use other inputs, because of the process of learning by using. In turn, this increases the profitability of producing p in all the subsequent years to a larger extent than that of producing other products, all else constant.

I summarized all these compound effects in the second chapter of this thesis in the following proposition:

Proposition 3 (*Effect of exogenous changes in the prices of inputs*): *A decline in the cost of an input k leads firms to add potential new products that use this input intensively. This effect is greater for firms with a higher firm-input-specific productivity to use k.*

3 Description of the database

This work uses a very detailed and comprehensive database for the Colombian manufacturing sector. It is often called EAM, which stands for its initials in Spanish (*"Encuesta Anual Manufacturera"*). I will use these initials throughout this document. The Administrative Department for Statistics of Colombia -DANE- started gathering information about the manufacturing firms of Colombia in 1950. Nowadays, it surveys each year every Colombian manufacturing firm with more than 10 employees and/or sales above an amount in Colombian pesos equivalent to 120k U.S. dollars in 2021, and also a representative sample of the smaller manufacturing firms.

The EAM contains several very important modules of information that played a key role in the process of constructing all the variables that were used

[6] Mathematically, this complementarity comes from the fact that the prices of inputs interact multiplicatively with the firm-input-specific productivities and the input-output coefficients. Please see the second chapter of this thesis for a detailed explanation of this.

in the regressions that I present in the subsequent sections of this paper. One of them is a module of products. It includes detailed information of every product sold by each manufacturing firm in each year. This information includes the unitary price of each product produced by each firm in each year, and also the quantity sold of each of them. It also includes both of these variables (unitary price and sold quantity) for exports. In addition, the EAM contains a module of inputs[7], which contains the purchased quantity and the unitary price of every input used by each firm in each year. This module also contains imported quantity and unitary price of imports for every imported input by each firm in each year.

The EAM contains also several firm-level modules. One of them includes detailed information of the employed workers. This module includes the separate number of workers by the type of work they perform. In some years this disaggregation was more detailed than in others. Namely, for some years it is possible to know separately the number of people in sales force, the number of managers, the number of administrative staff, the number of professional staff working in plant production and the number of plant operators. Unfortunately, the information of workers is less disaggregated for other years. As a consequence, I had to include all the workers in only two categories in every year, with the purpose of having consistency across years: production workers (which includes plant operators and professional staff working in plant production) and non-production workers (which includes managers, administrative and sales staff, and very importantly for this work, research and development staff, if these activities are performed by the firm). The EAM includes the total wages paid to workers in each of these categories.

The EAM includes some cost modules that contain relevant information that is not being used for this paper, but many researchers could be interested in using in the future. Namely, these costs modules include information of costs of energy and water, telecommunication services, taxes, interests and rent, among many other variables. In some years it includes complete modules about energy use, which intend to characterize the evolution of main energy sources of firms.

Very importantly for this research, the EAM includes several variables that allowed me to identify each firm and its production in many senses. To start, each firm and plant is uniquely identified with a numerical code that remains the same every year, as long as the firm is surveyed (which depends on the sales and number of employees, as explained before). This allowed me to trace each firm in time.

I used information of the EAM from 1992 to 2017, as information of years before 1992 has several problems and its organization and cleaning would have delayed this research substantially. As for the final year, 2017 was the last available year in the database when the bulk of this research was carried out. All the products and inputs are uniquely identified by a code of the International

[7] Please see the second chapter of this thesis for a detailed definition of "inputs". In short, this term makes reference to all the physical materials used in the productive process that are *directly* transformed into products. As capital and labour *are not* directly transformed into products, these are not included in this category of "inputs" in this thesis.

Standard Industrial Classification (ISIC) revision 2 adapted for Colombia from 1992 to 2000, by a code of the Central Product Classification (CPC) version 1.0 from 2001 to 2012, and by a code of the CPC version 2.0 from 2013 to 2017. In all cases, the EAM uses the most disaggregated levels available of each classification (8 digits for the ISIC rev. 2 and 9 digits for the CPC).

The fact that the EAM uses always the most disaggragated available codes to identify products and materials is very convenient for this research, as it ensures that a code truly identifies a specific product or material to the largest possible extent. However, this also prevented me from performing analysis that required tracing products or inputs across years with different classifications, as there do not exist concordance tables for these levels of disaggregation, but for much more aggregated categories of products. I could have used such levels of aggregation to gain traceability, but I would have lost specificity in the definition of products, which I valued more. As a consequence, I can only trace products within the periods 1992-2000, 2001-2012 and 2013-2017, which fortunately are not extremely short.

The modules of products and inputs are crucial for this work. Unfortunately, they are not publicly, as their free use could violate some Colombian laws that protect the privacy of information for some firms that might be identified even though firms and plants are anonymised. A good example is Reficar, the largest oil refinery of Colombia [8]. Reficar is located in the city of Cartagena. It produces a high share of the total amount of gasoline and diesel used in the country. Even though it is impossible to identify Reficar by its legal ID (as this latter is different from the identifier that was assigned to it in an anonymization process carried out by DANE), it is possible just to search for manufacturing plants in Cartagena in the business of refining, and there will be just one plant with sales as high as to correspond to Reficar. This would allow anyone to see sensible information such as unitary prices by product.

Because of the reasons explained in the last paragraph, the modules of products and inputs can only be accessed from an office located at DANE's main building. Unfortunately, it is not possible to guarantee that regressions that are left running by external researchers run in nighttime and for several days, as interruptions use to happen in the main system. For this reason, I had to use a random sample of firms, as using all firms with all their modules of products and inputs in all years would have implied processing times that exceed by far the time I was allowed to stay in their office. The pandemic of Covid-19 exacerbated this problem, as DANE's main building was closed for almost a year, and then reopened gradually at a very slow pace, with times as restrictive as only 8 hours per week for several months.

The total number of firms included in the EAM has been approximately 8.000 per year in the last two decades. As I am using data from 1992 to 2017, this means that my full firm-year-level database has in total approximately 208.000 observations (8.000 firms per year multiplied by 26 years). As each firm produces on average 3.5 products in a year and uses on average 25 mate-

[8] It has the capacity to refine 150.000 barrels of crude oil per day.

rials in a year, my firm-year-product-level database has in total approximately 728.000 observations, and my firm-year-material-level database has in total approximately 5.2 million observations.

The numbers shown at the end of the previous paragraph reveal the magnitude of the problem. As I will explain in the next section, I had to calculate for this work similarities between each firm and all its potential new products in each year [9] in terms of the use of inputs (or input mix), as such similarity is used as explanatory variable in all the regressions presented in the subsequent sections. This required the calculation of many dot products of two vectors of use of inputs for each product-firm combination for each year. This means that several procedures would have to have been applied to the 5.2 million observations if I had used my full database. I tried this, and unfortunately it exceeded the capacity of the computers available at DANE's office.

In order to bypass this problem, I took a stratified random sample of 2 percent of all the available firms in my full dataset, this is, 160 firms. Strata are defined by an interaction of two dimensions: quartiles of sales of firms in their initial year and number of produced products by firms in their initial year. This procedure guarantees that the sample includes firms that had different sizes in sales and different product scopes when they were born (or started complying with the sales and/or employment requirements to be included in the EAM). I selected these two variables because it is reasonable to expect that they have more influence on the process of product addition than the rest of firm-level variables for which I have information. More specifically, it is reasonable to expect that larger firms add more products to their product mixes, and also that firms that produced more products in their initial year add more products to their product mixes. Even though I do not analyze in depth the relationship of these variables with product addition, the possibility that this correlation exists explains their use as stratification variables.

The fact that each firm's probability of inclusion depends on the two aforementioned variables at their initial year and not on their averages across years prevents me from selecting a sample in which firms that started being small and grew both in sales and in product scope were over-represented. The number of firms selected from each strata is proportional to the share that each strata represents of the total number of firms. This ensures that all the strata are represented in the sample to an extent that is proportional to their importance in the total database of firms.

The initial sales and the initial number of products are correlated [10], which is not surprising, given that they were both chosen as stratification variables because of their possible relationship with product addition. This correlation reinforces the convenience of using them both as stratification variables, as this "bivariate" stratification guarantees that the firms *within* each quartile of sales

[9] Please see the second chapter of this thesis for a formal definition of potential new products. In short, this firm-year-specific set includes all the products that are feasible for the corresponding firm in the corresponding year, but are not produced by it in such year

[10] The correlation between the number of products in the initial year and the sales in that year is nearly 0.55.

are selected in such a way that the firms with different numbers of produts are represented proportionally to their importance in the respective quartile of sales.

This sampling procedure yields a firm-year-level dataset with approximately 7.000 observations, a firm-year-product-level dataset with approximately 14.000 observations and a firm-year-material-level dataset with approximately 100.000 observations.

In addition to information of Colombian manufacturing firms, this paper uses information of the tariffs imposed by Colombia to the imports of all products during the period 1992-2017. Namely, I use for each product in each year the average across all origin countries of the tariffs imposed to imports. This data was obtained from the World Bank's World Integrated Trade Solution (WITS). WITS includes several product classifications. I decided to use here information of products as defined by the Harmonized System (HS), as it is for this classification that I found the best possible table of concordances with the classifications of products provided by the Colombian office of statistics (International Standard Industrial Classification (ISIC) revision 2 adapted for Colombia, Central Product Classification (CPC) version 1.0 CPC version 2.0). As I will explain in detail in section 5, this data on tariffs is used to construct instrumental variables that are used in turn to analyze the possible existence of causal evidence in favor of the model proposed in the second chapter of this thesis.

4 Product addition, similarity in the input mix and skilled labour

This section explores possible empirical evidence in favor of the propositions 1 and 2 of section 2. I do this by estimating the parameters of econometric empirical models. I start by presenting the formal expression of the econometric model to be estimated. I then explain in detail how I calculated the dependent variable and the main regressor of this model. After this, I present and analyze key summary statistics of all these variables. Subsequently, I present the results of estimating the econometric model mentioned above under different specifications and for different spans (from $s = 1$ to $s = 5$). Finally, I present the results under a different way of defining the set of firms to be used in the regressions as an important robustness check.

Econometric model and expected results

Propositions 1 and 2 of section 2 predict that a specific product is more likely to be added by a firm in the future if it is more similar in terms of the input mix in the present time, and that this relationship between addition and similarity is stronger if such firm uses more skilled labour. However, a thorough analysis of the chapter 2 of this thesis reveals that the predicted relationship between product addition and similarity does not necessarily grow linearly as the amount of skilled labour increases [11].

[11] A detailed analysis of the second chapter of this thesis reveals that the extent to which

Given this, I use as regressors interactions of similarity with four dummies that are equal to one for firm-year observations that belong to quartiles one, two, three and four of the overall distribution of skilled labour (proxied here by the number of non-production workers[12]), respectively. In this way, I do not assume any functional form for the extent to which more skilled labour increases the size of the relationship between product addition and similarity, while still being able to analyze if this expected increase as a consequence of more skilled labour does occur. The parameter for each of these interactions measures how greater is the average relationship between product addition and similarity for the firms within the respective quartile of non-production workers than this relationship for the quartile whose interaction is excluded (in order to avoid perfect collinearity). As for the quartile whose interaction is excluded (which is always the lowest one), the relationship between product addition and similarity for the firms within it is given by the parameter for the similarity itself (which is also included in the model).

Quartiles of non-production workers themselves are also included as regressors because the proposition 2 of section 2 also predicts that the firms with more skilled labour add new products more easily, all else constant. In addition, this inclusion allows me to prevent the parameters for the interactions with similarity from capturing the separate effect of skilled labour (proxied by the number of non-production workers) on product addition, instead of its effect on the size of the relationship between product addition and similarity.

Formally, I estimate the parameters of the following linear model in order to analyze the empirical validity of the propositions 1 and 2 of section 2:

$$D_{fp,t+s} = \alpha + \sum_{q=2}^{4} \Gamma_q^s D_{qft} + \delta^s S_{fpt} + \sum_{q=2}^{4} \beta_q^s S_{fpt} * D_{qft} + \psi_t + \phi_f + \tau_p + \epsilon_{fp,t+s} \quad (1)$$

$D_{fp,t+s}$ represents here a dummy variable that equals one if a feasible product p that is not produced by the firm f in t is produced by f in $t+s$, and zero if it is not produced in $t+s$. S_{fpt} represents the similarity in t of the input mix needed to produce p with the input mix used by f. On the other hand, D_{qft} represents a dummy that equals one if the value of non-production workers employed by f in t (which is used here as a proxy for skilled labour) belongs to the quartile q of the total distribution of all observed values of non-production workers across all firms and years. ψ_t, ϕ_f and τ_p represent year, firm and product fixed effects, respectively. The four parameters β_q^s, the four parameters Γ_q^s and δ^s have a superscript s because they are specific for each time span. The summations across quartiles are from the second to the fourth quartile because the first quartile is excluded in both cases in order to prevent perfect collinearity.

the relationship between product addition and similarity grows as a firm uses more skilled labour is not characterized by a explicit function

[12] Please see the first chapter of this thesis for a complete discussion about the soundness of using this variable as a proxy of skilled labour

Given the proposition 1 stated in section 2, I expect δ^s to be positive for every $s \geq 1$ and $\delta^s + \beta^s_q$ to be also positive for every $q = 2...4$ and for every $s \geq 1$. In words, I expect that firms in all the quartiles of skilled labour are always more likely to add more similar products. Formally, this means that I expect the derivative of the expected value of $D_{fp,t+s}$ with respect to the similarity to be positive for every quartile of skilled labour.

Given that the proposition 2 in section 2 states that the advantage of the more similar products is greater in firms with more skilled labour, I expect $0 \leq \beta^s_2 \leq \beta^s_3 \leq \beta^s_4$. In words, I expect that increases in similarity are associated with higher increases in the probability of addition as I move to higher quartiles of skilled labour. As the proposition 1 in section 2 also states that less similar products (in terms of the input mix) are added as s increases, I expect the sum $\delta^s + \beta^s_q$ to fall as s increases for every $q = 2...4$, and I also expect δ^s to fall as s increases.

The proposition 2 in section 2 also states that firms with more skilled labour are on average more likely to add new products. Given this, I expect the sum $\Gamma_q + (\delta^s + \beta^s_q)E[S_{fpt}]$ to increase as q increases, for $q = 2...4$, and I also expect $\Gamma_2 + (\delta^s + \beta^s_2)E[S_{fpt}] > \delta^s E[S_{fpt}]$. In words, the latter condition guarantees that the expected value of $D_{fp,t+s}$ for the second quartile of skilled labour is greater than this expected value for the first quartile of skilled labour, for given a span s. The former condition implies that the same is true when the third quartile is compared to the second and first quartiles, and when the fourth quartile is compared to the third, second and first quartiles. As I show below in Table 1, the empirical estimator for $E[S_{fpt}]$ (this is, the sample mean of S_{fpt}) is 0.48. Therefore, I expect $\Gamma_q + 0.48(\delta^s + \beta^s_q)$ to increase as q increases and $\Gamma_2 + 0.48(\delta^s + \beta^s_2) > 0.48\delta^s$, for $q = 2...4$ and for every $s \geq 1$.

Finally, the error term $\epsilon_{fp,t+s}$ is assumed to be normally distributed with mean zero for every $s \geq 1$. These errors are clustered by firm. In other words, they are assumed to be independent across firms, but are allowed to be correlated with each other across time and products within each firm, and none theoretical structure is imposed for such correlation. Instead, such structure is directly estimated from the data.

Definition and construction of variables

The set of feasible products for a firm f (which is needed to establish the domain of the variable $D_{fp,t+s}$ in each case) is defined in a completely empirical way. Namely, it includes all the products that were ever produced by any firm simultaneously with any of the products produced by f in any year.

The variable S_{fpt} represents the similarity in terms of the use of materials between the firm f and the product p in t. It is calculated as a dot product of 2 vectors of expenditure shares on all the available inputs. One of them (x_{ft}) contains the observed shares for a firm f in t. The other (x_{pt}) contains the average of shares across all the firms in the sample that produce a product p in t. This dot product yields a number between zero and one, as it is normalized by dividing its value by the product of the total variability of both vectors. A value of one would indicate that the input mix needed to produce p in t is identical to the input mix used by f in t. In contrast, a value of zero would indicate that they are completely different. Formally, I calculate S_{fpt} by using the following formula, as in Boehm et al. (2019):

$$S_{fp}^{t} = \frac{\sum_{k=1}^{K} x_{fkt} x_{pkt}}{\left[\left(\sum_{k=1}^{K} x_{fkt}^{2}\right)\left(\sum_{k=1}^{K} x_{pkt}^{2}\right)\right]^{1/2}} \tag{2}$$

Summary statistics

The table 1 below shows different summary statistics of the variables that are used in this section to estimate the model described in (1) for different time spans. Some variables are observed at the firm-product-year level, whereas others are observed just at the firm-year level. In all the cases I used the sample described in detail in section 3.

The variables named *D(future production)* are dummy variables that equal one if a feasible product is produced and zero otherwise [13]. They can indicate in a year t the production of a given product in the current year or k years ahead, for $k = 2...5$.

For each time span I present the summary statistics for *D(future production)* in two different cases. In the first case (*"All products"*) the summary statistics are calculated by using the *future production* dummies of *all* the products in the whole feasible set of the corresponding firm. Formally, the means estimate in this case the *unconditional* probabilities of future production. In the second case (*"Products not being produced in t"*) the summary statistics are calculated by using just the *future production* dummies of those products that belong to the firm's feasible set *and are not produced by the firm in t.* Formally, the means estimate in this case the *conditional* probabilities of future production.

The first pane (textit "All firms") shows the summary statistics when all the sampled firms are included. The other four panes show the statistics when only the firms that belong to the $q - th$ quartile of non-production workers are included, for $q = 1...4$. This disaggregation across quartiles of non-production workers matters because this variable is used here as a proxy of skilled labour, and the propositions presented in section 3 state that the amount of skilled labour used by a firm determines to some extent its capacity to add new products to its product mix in the future.

The fact that the median is zero for all the cases indicates that most feasible products are not added by the firms to their product mixes, neither in the present time nor in the future (up to 5 years ahead). As expected, the

[13] As explained above, the set of feasible products of a firm f includes all the products that were ever produced by any firm simultaneously with any of the products produced by f in any year.

conditional probabilities are much lower than the unconditional ones. The conditional probabilities truly quantify the probabilities of addition of potential *new* products in the future, as they correspond only to products that are *not* produced in t. On average, between 3% and 7% of the potential new products are added in the future, depending on the time span and the amount of non-production workers. Very importantly, the probability of addition is higher for the highest quartile than for all the other quartiles for all the time spans, and the probability of addition is higher for the second quartile than for the first quartile for all the time spans.

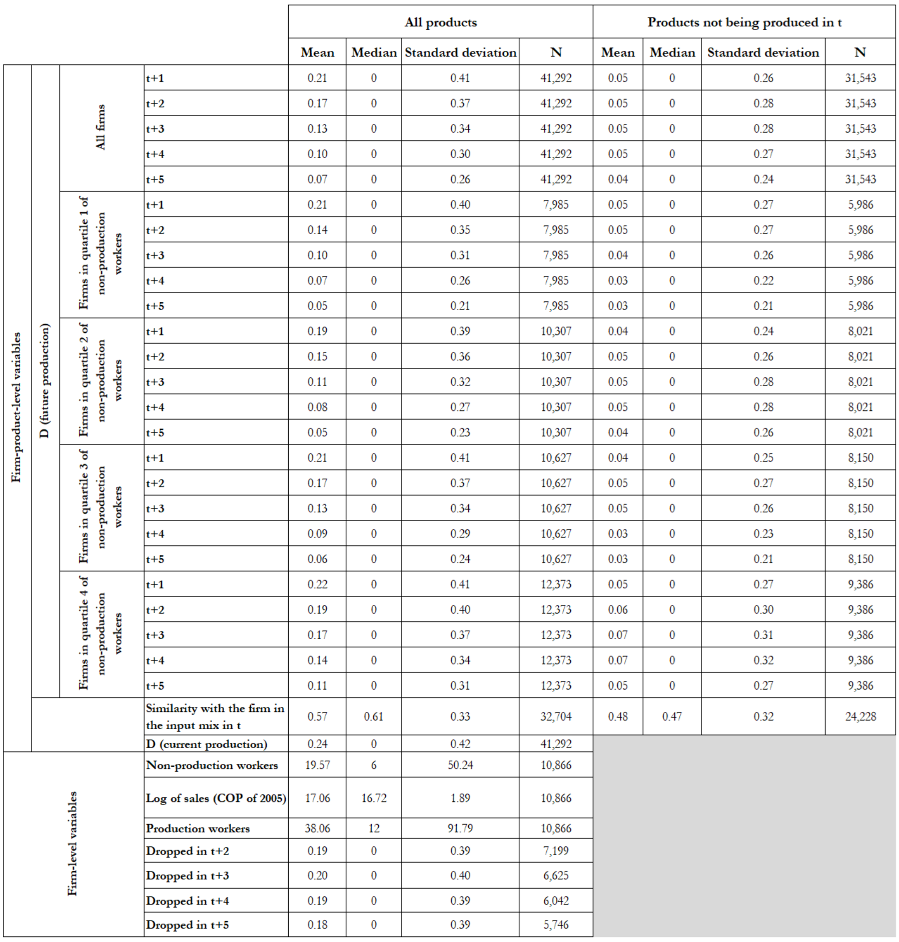
The row below the quartiles (*"Similarity with the firm in the input mix in t"*) shows that the feasible products that are effectively produced by a firm in t are on average more similar to each other than to those feasible products that are *not* produced by this firm in t. In other words, the average similarity of the produced products with the rest of produced products is higher than the average similarity of the non-produced products with the produced products. This is consistent with the propositions presented in Section 2.

As for the statistics for the firm-level variables, it is to note that firms use on average approximately two production workers per each non-production worker. It is also to note that the probability that a firm that exists in t drops at least one product from its product mix in $t+k$ does not change much across different values of k (for $k = 2...5$).

Results

In order to obtain consistent estimators for the parameters in equation (1), I run a regression with ordinary least squares with information from the EAM (see chapter 3 for details of this database). Namely, I use the module of the EAM that includes information of prices and quantities of all the inputs used by all the firms in every year to calculate S_{fpt} for all the products in the feasible set of every firm in the sample. I use the module of the EAM that includes information of prices and quantities of all the products sold by all firms in every year to define the feasible set of each firm and to construct the dummy variables $D_{fp,t+s}$. The number of non-production workers was directly taken from the EAM and used in the regression.

I start by presenting and interpreting the results of estimating the model (1) for $s = 1$, and then I move to longer spans. This because it is convenient to analyze initially the short-term relationship between product addition and the interaction of similarity with non-production workers, and then analyze separately such relation in longer terms. This distinction is relevant in the empirical work presented here because all the regressors (similarity, dummies for quartiles of non-production workers and their interactions) are all defined in the present time t, and their relationships with future product addition might be easier to identify in the short term than in the long term, when other excluded dynamic factors might affect the pattern of product addition to a larger extent than in the short term. I present also the results of regressions in which only the quartiles of non-production workers and the similarity are included separately as regressors.



				All products				Products not being produced in t			
				Mean	Median	Standard deviation	N	Mean	Median	Standard deviation	N
Firm-product-level variables	D (future production)	All firms	t+1	0.21	0	0.41	41,292	0.05	0	0.26	31,543
			t+2	0.17	0	0.37	41,292	0.05	0	0.28	31,543
			t+3	0.13	0	0.34	41,292	0.05	0	0.28	31,543
			t+4	0.10	0	0.30	41,292	0.05	0	0.27	31,543
			t+5	0.07	0	0.26	41,292	0.04	0	0.24	31,543
		Firms in quartile 1 of non-production workers	t+1	0.21	0	0.40	7,985	0.05	0	0.27	5,986
			t+2	0.14	0	0.35	7,985	0.05	0	0.27	5,986
			t+3	0.10	0	0.31	7,985	0.04	0	0.26	5,986
			t+4	0.07	0	0.26	7,985	0.03	0	0.22	5,986
			t+5	0.05	0	0.21	7,985	0.03	0	0.21	5,986
		Firms in quartile 2 of non-production workers	t+1	0.19	0	0.39	10,307	0.04	0	0.24	8,021
			t+2	0.15	0	0.36	10,307	0.05	0	0.26	8,021
			t+3	0.11	0	0.32	10,307	0.05	0	0.28	8,021
			t+4	0.08	0	0.27	10,307	0.05	0	0.28	8,021
			t+5	0.05	0	0.23	10,307	0.04	0	0.26	8,021
		Firms in quartile 3 of non-production workers	t+1	0.21	0	0.41	10,627	0.04	0	0.25	8,150
			t+2	0.17	0	0.37	10,627	0.05	0	0.27	8,150
			t+3	0.13	0	0.34	10,627	0.05	0	0.26	8,150
			t+4	0.09	0	0.29	10,627	0.03	0	0.23	8,150
			t+5	0.06	0	0.24	10,627	0.03	0	0.21	8,150
		Firms in quartile 4 of non-production workers	t+1	0.22	0	0.41	12,373	0.05	0	0.27	9,386
			t+2	0.19	0	0.40	12,373	0.06	0	0.30	9,386
			t+3	0.17	0	0.37	12,373	0.07	0	0.31	9,386
			t+4	0.14	0	0.34	12,373	0.07	0	0.32	9,386
			t+5	0.11	0	0.31	12,373	0.05	0	0.27	9,386
			Similarity with the firm in the input mix in t	0.57	0.61	0.33	32,704	0.48	0.47	0.32	24,228
			D (current production)	0.24	0	0.42	41,292				
Firm-level variables			Non-production workers	19.57	6	50.24	10,866				
			Log of sales (COP of 2005)	17.06	16.72	1.89	10,866				
			Production workers	38.06	12	91.79	10,866				
			Dropped in t+2	0.19	0	0.39	7,199				
			Dropped in t+3	0.20	0	0.40	6,625				
			Dropped in t+4	0.19	0	0.39	6,042				
			Dropped in t+5	0.18	0	0.39	5,746				

Table 1. Summary statistics of relevant firm-product-level and firm-level variables.

Table 1 presents the results of estimating the model (1) for $s = 1$ by ordinary least squares as described before. The third column shows the results when all the regressors in the equation (1) are included. The most important finding is that $\delta^1 + \beta^1_q$ is positive as expected for $q = 2$ and $q = 4$. Moreover, the estimator for the parameter β_4 (that is, for the highest quartile of non-production workers) is larger than the estimators for β_2 and β_3. This means that the correlation between similarity and product addition is unambiguously larger for the firms in the highest quartile non-production workers than for the firms in all the other quartiles. The fact that the estimator for β_2 is positive and statistically significant means that the firms in the second quartile of non-production workers have a higher correlation between product addition and similarity than the firms in the first quartile of this variable.

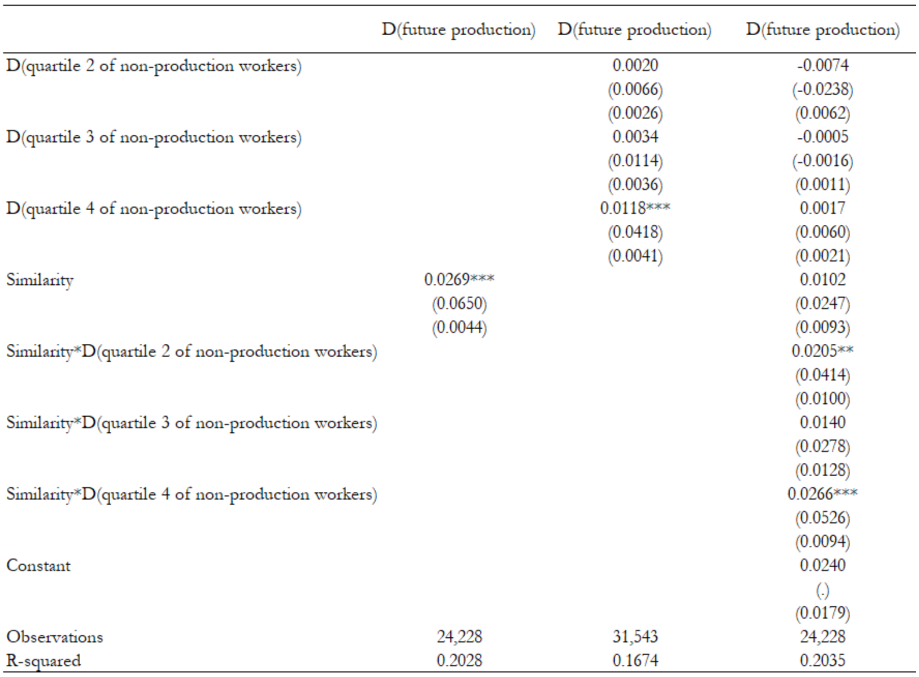
The only fact in Table 2 that contradicts the proposition 1 of section 1 so far is that the correlation between product addition and similarity is not larger for firms in the third quartile than for those in the second quartile. I ran an alternative specification of the model (1) for $s = 1$ in which I included the logarithm of sales as a regressor, in order to avoid a possible omitted variable bias (as the number of non-production workers of a firm is positively correlated with its size, proxied here by its sales). However, the main conclusions summarized so far remain in this case.

Very importantly, the consistency of my empirical results with the theoretical prediction that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$ increases when I include the potential cost of production as a regressor. I will show this result in the section 5 of this paper. The inclusion of this variable might seem redundant with the inclusion of similarity, as it is possible that more similar products are added more easily precisely *because* they are cheaper (and therefore more profitable) for firms. Suppose that a firm f is very good at using glass in t. This makes its cost of producing windows (w) in $t+1$ low. As a consequence, f adds w to its product mix in $t+1$. f is also likely to produce glass-intensive products in t. In this example, the similarity between w and $f's$ production in t in terms of the input mix is high, as both mixes are intensive in glass. In addition, the potential cost of w is low in $t+1$. Both facts (the high similarity of w and its low potential cost) come from the fact that f is good at using glass. Given this, the inclusion of both variables in a regression might be redundant.

However, it is possible that the empirical similarity depends also on other variables different from the drivers of the potential cost. Firms might be forced to use in reality inputs in certain combinations because of physical restrictions in the availability of certain inputs. There may be also technical innovations that lead firms to use temporarily or permanently certain inputs for whose use they do not necessarily have high firm-input-specific productivities. If this were the case, two implications would arise. Firstly, including both the potential cost and the similarity in a regression would not be completely redundant, as they reflect similar but different drivers. Secondly, the similarity might be less correlated than the potential cost with product addition. This because the similarity might indicate just (i) the input mix that a firm *has* to use to produce a product, whereas the potential cost would measure (ii) its intrinsic capacity to produce it cheaply. It is reasonable to expect that (ii) is more correlated with product addition than (i), as long as product addition is more determined by firms' perception of their capacities than by their actual use of inputs.

Given these reasons, I will explore in the section 5 of this paper what happens when both the similarity and the potential cost of production are included. The main conclusion will be that in this case there is stronger evidence that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$, as expected. I interpret this as suggestive evidence of a possible omitted variable bias in the specification (1). Please see the section 5 for a detailed discussion of these results.

As expected, the sum $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ is greater for the fourth quartile than for all the lower quartiles, and $\Gamma_2 + 0.48(\delta^1 + \beta_2^1) > 0.48\delta^1$. However, the same cannot be concluded for the third quartile when compared to the first and second ones. Therefore, the evidence about the hypothesis that firms with more non-production workers are more likely to add new products is not conclusive for $s = 1$.



	D(future production)	D(future production)	D(future production)
D(quartile 2 of non-production workers)		0.0020	-0.0074
		(0.0066)	(-0.0238)
		(0.0026)	(0.0062)
D(quartile 3 of non-production workers)		0.0034	-0.0005
		(0.0114)	(-0.0016)
		(0.0036)	(0.0011)
D(quartile 4 of non-production workers)		0.0118***	0.0017
		(0.0418)	(0.0060)
		(0.0041)	(0.0021)
Similarity	0.0269***		0.0102
	(0.0650)		(0.0247)
	(0.0044)		(0.0093)
Similarity*D(quartile 2 of non-production workers)			0.0205**
			(0.0414)
			(0.0100)
Similarity*D(quartile 3 of non-production workers)			0.0140
			(0.0278)
			(0.0128)
Similarity*D(quartile 4 of non-production workers)			0.0266***
			(0.0526)
			(0.0094)
Constant			0.0240
			(.)
			(0.0179)
Observations	24,228	31,543	24,228
R-squared	0.2028	0.1674	0.2035

Table 2. Results of regression of product addition as a function of similarity in the input mix interacted with dummies for quartiles of non-production workers. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

The numbers below the estimators in Table 1 are the beta coefficients that correspond to each one of them. The beta coefficient for a specific regressor is calculated from its estimator and the standard deviations of both that regressor and the dependent variable. Formally, a beta coefficient for a variable x measures how many standard deviations the dependent variable grows or falls on average when the corresponding regressor grows one standard deviation. This allows an easier economic interpretation of the size of the correlations between the regressors and the dependent variable. The numbers two lines below the estimators are the estimated standard errors. I report beta coefficients and standard errors in this same order in the tables 3 and 4 below. [14].

Summarizing, I have found so far mixed evidence of the possible validity of propositions 1 and 2 of chapter 2. More specifically, the evidence is in favor of these propositions when the firms with the highest values of skilled labour are compared to the rest of firms. Firstly, the firms with the highest values of skilled labour (proxied by the number of non-production workers) exhibit

[14] In the table 5 I report just the p-values because in that case the size of the correlations is not relevant, whereas it will be more important to analyze if the null hypothesis of non-significance of the instrumented variables used to test for causality can or cannot be rejected.

a higher correlation between similarity and product addition than the rest of firms. Similarly, firms in the second quartile of non-production workers exhibit a higher correlation between these two variables than the one observed for the firms in the first quartile. Finally, the expected value of the dummy for product addition one year ahead $D_{fp,t+1}$ is on average higher for the firms in the fourth quartile of the number of non-production workers than for the rest of the firms, as expected.

Let us suppose that there exists a firm $f1$ with a level of skilled labour (non-production workers) that belongs to the second quartile of this variable in t. According to Table 1, an increase in t of one standard deviation of the similarity between a product p and $f1$ in terms of the input mix makes $f1$ 4.1% more likely to add p than a firm in the bottom quartile of skilled labour (non-production workers). In contrast, if other firm $f2$'s skilled labour (number of non-production workers) belongs to the fourth quartile of this variable in t, an increase in t of one standard deviation of the similarity between a product p and $f2$ in terms of the input mix makes $f2$ 5.3% more likely to add p than a firm in the bottom quartile of skilled labour (non-production workers). On the other hand, the average probability of addition for $f1$ of a product with a similarity of (say) 0.4 is $0.024 + 0.0102 * 0.4 + 0.0205 * 0.4 = 3.6\%$. In contrast, the average probability of addition for $f2$ of a product with a similarity of (say) 0.4 is $0.024 + 0.0102 * 0.4 + 0.0266 * 0.4 = 3.9\%$.

Given that the results found for $s = 1$ are partially in favor of the propositions 1 and 2 of section 2, it is reasonable to expect the same for $s > 1$. In order to analyze the possible empirical validity of these propositions for $s > 1$, I ran four regressions with ordinary least squares that correspond to the model (1) for $s = 2...5$. Table 3 shows the results.

A first fact that stands out from Table 3 is that the number of observations falls as the time span becomes longer in most of cases. This happens because of two factors. Firstly, some firms in my sample disappear as the time passes, so the number of firms that remain in the sample for s years falls as s increases (that is, as the span of analysis becomes longer). Secondly, the number of years included in the regressions falls as s becomes larger, as in any regression that includes the lead of at least one variable. Namely, one year is lost every time s increases one unit (year). As the total number of observations in the regression for a specific s equals the total number of years minus s multiplied by the number of firms that survive s years after each year in the sample multiplied by the number of products in the feasible set of each firm (which does not change over time for a firm), the observed decreases in the total number of observations is exclusively attributable to the two factors mentioned above.

Another fact that must be mentioned about the regressions in Table 3 is that the fitting power of the models is larger for longer spans than for $s = 1$. The coefficient of determination ($R2$) is greater than 30 percent in three of the four cases, well above the $R2$ of 20% for $s = 1$. This might be due to several factors. Firstly, the inclusion of the variable *"Dropped product(s) in t+s"* might be increasing the proportion of total variability in product addition predicted by the model. This variable will be explained in detail later in this section. On the other hand, $R2$ might also be higher because the firms included in the regressions exhibit less variability in their patterns of product addition as s grows from 1 to higher numbers.

	D(future production in t+2)	D(future production in t+3)	D(future production in t+4)	D(future production in t+5)
D(quartile 2 of non-production workers)	0.0074	0.0193***	0.0243***	0.0298***
	(0.0224)	(0.0590)	(0.0781)	(0.1022)
	(0.0069)	(0.0068)	(0.0073)	(0.0101)
D(quartile 3 of non-production workers)	0.0041	0.0111	0.0106	0.0133
	(0.0128)	(0.0344)	(0.0353)	(0.0477)
	(0.0076)	(0.0075)	(0.0077)	(0.0094)
D(quartile 4 of non-production workers)	0.0076	0.0151*	0.0176*	0.0141
	(0.0256)	(0.0516)	(0.0640)	(0.0553)
	(0.0080)	(0.0091)	(0.0092)	(0.0104)
Similarity	0.0175*	0.0329***	0.0301***	0.0226*
	(0.0395)	(0.0744)	(0.0720)	(0.0583)
	(0.0102)	(0.0108)	(0.0108)	(0.0134)
Similarity*D(quartile 2 of non-production workers)	0.0086	-0.0242**	-0.0258**	-0.0244*
	(0.0158)	(-0.0461)	(-0.0504)	(-0.0509)
	(0.0111)	(0.0118)	(0.0115)	(0.0136)
Similarity*D(quartile 3 of non-production workers)	0.0045	-0.0259**	-0.0182	-0.0143
	(0.0082)	(-0.0467)	(-0.0350)	(-0.0297)
	(0.0130)	(0.0128)	(0.0120)	(0.0141)
Similarity*D(quartile 4 of non-production workers)	0.0169	-0.0166	-0.0092	-0.0063
	(0.0326)	(-0.0318)	(-0.0184)	(-0.0136)
	(0.0123)	(0.0126)	(0.0127)	(0.0146)
dropped	0.0220***	0.0201***	0.0226***	0.0142***
	(0.0586)	(0.0563)	(0.0660)	(0.0440)
	(0.0044)	(0.0039)	(0.0041)	(0.0034)
Constant	0.0335	0.0606*	0.0895**	0.1042***
	(.)	(.)	(.)	(.)
	(0.0287)	(0.0318)	(0.0360)	(0.0365)
Observations	17,622	16,218	14,737	14,048
R-squared	0.2932	0.3021	0.3114	0.3096

Table 3. Results of regressions of product addition in future years from $t+2$ to $t+5$ as a function of similarity in the input mix interacted with dummies for quartiles of non-production workers. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

As expected, δ^s is positive and statistically significant for every $s = 2...5$. Also as expected, $\delta^s + \beta_q^s$ is positive for all the non-excluded quartiles of non-production workers (that is, for all q from 2 to 4) and for every $s = 2...5$, with the only exception of the firms in the lowest quartile of non-production workers in $t + 5$. These facts imply that the derivative of $D_{fp,t+s}$ with respect to similarity is positive for all quartiles and spans, except for the firms in the lowest quartile of non-production workers in $t + 5$. However, I cannot conclude that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$ for any s. Therefore, the aforementioned derivative does not grow as I move to higher quartiles of non-production workers.

The evidence does not allow me to conclude in general that the sum $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ grows as q grows. In other words, there is not evidence that the firms with more non-production workers are more likely to add new potential products for $s = 2...5$.

As for the evolution of estimators over time, I cannot conclude from Table 3 that δ^s and $\beta_q^s + \delta^s$ fall over time in all cases, as expected. The evidence is mixed in this case. To start, δ^s grows from $s = 2$ to $s = 3$, but then falls monotonically from $s = 3$ to $s = 5$. As for $\beta_q^s + \delta^s$, it falls monotonically from $s = 2$ to $s = 5$ for quartile 2. For quartiles 3 and 4 it falls from $s = 2$ to $s = 3$, then grows from $s = 3$ to $s = 4$, and falls again from $k = 4$ to $k = 5$.

Summarizing, I found that the derivative of product addition with respect to similarity is positive on average for firms of all the quartiles of non-production workers in all spans from $s = 2$ to $s = 5$. This is consistent with the expected results. However, I did not find conclusive evidence that this derivative is in general greater in firms with more non-production workers. As for the dynamics of this derivative, I could not establish conclusive evidence that it falls over time, as expected.

The variable "Dropped" is a dummy equal to one if the firm drops a product in the respective year $t + s$, and zero otherwise. It intends to capture the possibility (not explored in the model of the second chapter of this thesis) that firms has scarce factors of production that are necessary for the production of several products and that cannot be used simultaneously for the production of several products to a full extent, and that therefore they need to drop some existing products to add others. The possible existence of these factors was theorized by Sutton (2012), who considers it a key determinant of product scope of countries. This variable is positive and statistically significant at a level of significance of 1 % for every $s = 2...5$, which suggests the possible presence of factors of production with the properties mentioned above. A deep exploration of their role and nature at the firm level remains as a pending and crucial task for future works.

It is worth to mention that the analyses of significance presented so far might be influenced by the number of firms that I am using. Namely, my sampling strategy implies that I am using only 160 firms, which in turn implies that I use just 160 clusters to estimate the standard errors of the estimators. A low number of clusters may lead in general to incorrect rejections of null hypotheses that true parameters are equal to zero[15] (this is, this can lead to claiming statistical significance in cases in which it does not exist). However, 160 is a larger number than the minimum number of clusters identified in the literature for the statistical inference to be still valid (which is about 50) [16].

Robustness check: keeping the same sample of firms across years

It is possible that the changes of estimators across different spans for a given quartile of non-production workers described in the previous section be partially attributable to changes in the sample. As I explained before, some firms disappear from the sample as I move to longer spans (higher values of s) and the sample size falls mechanically as I do so. These facts might explain to some extent that changes of estimators across different spans differ from the expected ones. In order to explore this possibility, Table 4 shows results for the same regressions of Table 3, with the only difference that the regressions in table 4 are all run with the same sample. Namely, I ran all the regressions of Table 4 using only the firms that survived continuously from t to $t+5$ and using in every case only information of the years that can be used for all the four regressions.

The main conclusion from Table 4 is that the results after restricting the sample of firms to avoid a possible attrition bias are not very different from the ones that I found with the whole sample of firms. In this restricted case, $\delta^s > 0$ and $\delta^s + \beta_q^s > 0$ for $s = 3$, $s = 4$ and $s = 5$, and for all quartiles of non-production workers. In words, the derivative of $D_{fp,t+s}$ with respect to similarity is positive for all quartiles for all those three spans (3, 4 and 5 years ahead). This is not very different from the results in Table 3. Just as for Table 3, I cannot conclude from table 4 that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$ for any s. Therefore, the aforementioned derivative does not grow either in this case as I move to higher quartiles of non-production workers. Once again, I cannot conclude from Table 4 that $\Gamma_q + 0.48(\delta^s + \beta_q^s)$ grows as q grows in general for $s = 2...5$. Finally, I cannot conclude from Table 4 that δ^s and $\beta_q^s + \delta^s$ fall over time in all cases, as expected. This conclusion is identical to the one I presented for the case with the whole sample of firms.

In addition to the possible problems induced by the exit of firms over time, I had the concern that the variability of the dependent variables of the regressions whose results are shown in tables 2 and 3 might be too low. This because the dummies for future product addition have too many zeros, as the feasible sets include many more products than the ones that are in fact produced by the firms. This problem might have two negative implications. Firstly, it might reduce the statistical power of the models estimated here. Secondly, the variance of the errors might be too low, leading to incorrect inferences about the statistical significance of the estimators. In order to analyze if this is the case, I ran the regressions of table 3 with a more restrictive definition of the feasible set.

I defined the feasible set in an alternative way in which this set contains for a firm f all the products that f has ever produced and all the products with

[15] Please read Cameron et al. (2008) for details.

[16] Please read Angrist and Pischke (2009) for a detailed explanation.

a probability of joint production with any product ever produced by f above the median of the distribution of all the probabilities of joint production (unlike the original definition, in which this set includes *all* the products that were ever jointly produced with any of the products ever produced by f). This reduces the number of zeros in the dependent variable (the dummy for product addition), and prevents the problem of too small variability explained before. However, I did not observe important changes in this case with respect to the results shown in table 3.

	D(future production in t+2)	D(future production in t+3)	D(future production in t+4)	D(future production in t+5)
D(quartile 2 of non-production workers)	-0.0021	0.0138	0.0238***	0.0362***
	(-0.0062)	(0.0399)	(0.0695)	(0.1108)
	(0.0099)	(0.0099)	(0.0091)	(0.0109)
D(quartile 3 of non-production workers)	-0.0093	0.0034	0.0113	0.0183*
	(-0.0302)	(0.0102)	(0.0353)	(0.0603)
	(0.0100)	(0.0095)	(0.0092)	(0.0103)
D(quartile 4 of non-production workers)	-0.0044	0.0066	0.0192*	0.0183
	(-0.0160)	(0.0225)	(0.0675)	(0.0679)
	(0.0107)	(0.0113)	(0.0107)	(0.0115)
Similarity	0.0188	0.0337**	0.0283**	0.0342**
	(0.0435)	(0.0738)	(0.0635)	(0.0809)
	(0.0146)	(0.0148)	(0.0136)	(0.0155)
Similarity*D(quartile 2 of non-production workers)	0.0075	-0.0164	-0.0148	-0.0310**
	(0.0133)	(-0.0282)	(-0.0250)	(-0.0560)
	(0.0162)	(0.0167)	(0.0149)	(0.0151)
Similarity*D(quartile 3 of non-production workers)	0.0064	-0.0235	-0.0068	-0.0157
	(0.0118)	(-0.0400)	(-0.0121)	(-0.0295)
	(0.0169)	(0.0160)	(0.0143)	(0.0151)
Similarity*D(quartile 4 of non-production workers)	0.0109	-0.0149	0.0002	-0.0098
	(0.0219)	(-0.0280)	(0.0004)	(-0.0201)
	(0.0164)	(0.0164)	(0.0153)	(0.0164)
dropped	0.0219***	0.0257***	0.0221***	0.0156***
	(0.0597)	(0.0691)	(0.0598)	(0.0449)
	(0.0054)	(0.0054)	(0.0050)	(0.0043)
Constant	0.0573	0.0804**	0.1018**	0.1045**
	(.)	(.)	(.)	(.)
	(0.0372)	(0.0383)	(0.0419)	(0.0417)
Observations	11,625	11,625	11,625	11,625
R-squared	0.2981	0.3292	0.3669	0.3601

Table 4. Results of regressions of product addition in future years from $t+2$ to $t+5$ as a function of similarity in the input mix interacted with dummies of quartiles for non-production workers keeping the same sample. Beta coefficients in parentheses one line below the estimators. Estimated standard errors two lines below the estimators.

Discussion

The results presented in this section are mostly indicative of the existence of a positive derivative of product addition with respect to similarity that remains over time, although it does not change over time in the way I expected. In simpler words, firms keep on adding to their product mixes even in the long term products whose production required in t input mixes that were more similar to the input mixes used by them in that year. This finding is important because this persistence is consistent with the main and distinctive feature of the model that I proposed in the second chapter of this thesis, which is the persistence of firm-input-specific productivities that cause persistent patterns of product addition.

On the other hand, the empirical results presented here are not fully consistent with the expected result that the derivative of product addition with respect to similarity is greater in firms with more non-production workers. However, I did find that the firms with the highest values of non-production workers exhibit a larger derivative than the rest of firms in the short term (in $t+1$).

The results presented here do not suggest unambiguously that the derivative of product addition with respect to similarity falls over time for every quartile of non-production workers, as expected. There are several possible explanations for this. One that is consistent with the model of the second chapter of this thesis is that firm-input-specific productivities might depreciate over time. If this were the case, such productivities might grow or fall over time, depending on the size of the depreciation and on the scale of the process of learning by using. Depreciation might dominate in some periods, leading firms to add more (and not less) similar products, as they must rely to a larger extent in this case on what they already know how to do well. I do not explore this or other possible explanations in this paper.

5 The effect of cheaper inputs on product addition

In this section I analyze if the exogenous changes in the prices of inputs have in reality the firm-product-specific effects predicted by the theoretical model presented in the second chapter of this thesis. Very importantly, I explain here why the granularity of this prediction allows me to claim that finding possible validity for it would constitute causal evidence in favor of the theoretical model presented in the second chapter of this thesis.

The Proposition 3 in section 2 predicts what is expected to happen to product addition if the price of an input used by the firms falls. It states that this reduction has firm-product-specific effects on the profitability of product addition. More specifically, this effect depends in each case on two differential elements: (i) it is larger if the respective product is intensive in the input whose price falls, and (ii) it is larger if the respective firm has higher firm-input-specific productivity for the use of the inputs whose price falls.

The ideal way to analyze empirically the possible validity of the firm-product-specific prediction contained in the proposition 3 would be to construct a variable that reflects the two elements (i) and (ii) mentioned above, and analyze its possible correlation with the phenomenon of product addition. This variable should ideally have two characteristics. Firstly, it should be possible to calculate it in reality, as its observed values are needed if I want to use it in an empirical analysis. Secondly, it should *capture* the two differential elements (i) and (ii) above. By "capture" I mean that this variable should respond to the changes in the prices of inputs in directions and magnitudes that reflect the elements (i) and (ii) above.

Fortunately, the theoretical model in the second chapter of this thesis offers an ideal candidate for the variable mentioned above. The expression (6) of the second chapter of this thesis is the conditional cost function of a firm f to produce a quantity Q_{fp} of a product p in a year t. Formally, this cost has the following functional form:

$$\mathrm{C}_{fpt} = \left[\sum_k \phi_{fkt}^{\sigma} \phi_{pk}^{\sigma} q_{fkt}^{1-\sigma}\right]^{1/1-\sigma} \frac{Q_{fp}^{1/\theta}}{g_p\left(\overline{K_{ft}}, \overline{L_{ft}}\right)^{\beta}/\theta}$$

(3)

C_{fpt} has the key property that it determines the phenomenon of product addition (this is why it is the first result presented in the second chapter of this thesis). Intuitively, if a firm f can produce a product p in a year t at a lower cost (this is, if C_{fpt} above is lower), then p is more profitable for f and it is easier for f to add it to its product mix in the future. In short, if C_{fpt} is higher, then the profitability of adding p for f in t is lower.

Very importantly, C_{fpt} is lower *for every produced quantity* if the term $\left[\sum_k \phi_{fkt}^{\sigma_p} \phi_{pk}^{\sigma_p} q_{fkt}^{1-\sigma_p}\right]^{1/1-\sigma_p}$ is lower. In other words, this component scales up or down the minimum cost of production for any produced quantity. Given its relevance for the analysis below, I will name this term *CES cost index* (CCI) hereinafter, as I will use it repeatedly later. Formally:

$$\mathrm{CCI}_{fpt} \equiv \left[\sum_k \phi_{fkt}^{\sigma} \phi_{pk}^{\sigma} q_{fkt}^{1-\sigma}\right]^{1/1-\sigma}$$

(4)

What makes CCI_{fpt} special for the purposes of this paper is that it captures formally the intuition explained before for the facts that the firm-product-specific effect of the reduction in the price of an input is (i) larger if the respective product is intensive in the input whose price falls, and (ii) larger if the respective firm has a higher firm-input-specific productivity for the use of the input whose price falls. In other words, CCI_{fpt} has the desired property of capturing (i) and (ii). The part (i) is captured by the fact that the price of each input q_{fkt} is multiplied by ϕ_{pkt} (the input-output coefficient), which amplifies the effect of falls in prices of inputs with high ϕ_{pk}. The part (ii) is captured by the fact that the price of each input q_{fkt} is also multiplied by ϕ_{fkt} (the firm-input-specific productivity), which amplifies the effect of falls in prices of inputs with high ϕ_{fkt}. Please notice that the fact that both effects are explained by the triple product $\phi_{fkt}\phi_{pk}q_{fkt}$ (each with its respective exponential) implies that the two differential factors (i) and (ii) amplify each other. This latter fact is not explored theoretically nor empirically in this thesis, but it might be explored in the future.

So far, I have explained why CCI_{fpt} has the key characteristic of capturing properly (i) and (ii). In short, CCI_{fpt} determines the cost of production (which in turn determines the phenomenon of product addition), and it changes when the prices of inputs fall in firm-product-specific magnitudes that reflect the differential elements (i) and (ii). However, I have not demonstrated yet that CCI_{fpt} has the other desired property of being possible to calculate. Fortunately, CCI_{fpt} can be calculated indeed, as I will show later in this section. Before explaining how, I will explain how it is used, what are the expected results when using it, and what econometric problems might arise because of its use.

Econometric model and expected results

Given the very important characteristics of CCI_{fpt}, I use it as a regressor for the probability of product addition. More specifically, I include it as an additional regressor in the first econometric model of the previous section. Namely, I estimate here the parameters of the following expression:

$$\mathrm{D}_{fp,t+1} = \alpha + \sum_{q=1}^{4} \beta_q^s S_{fpt} * D_{qft} + \sum_{q=1}^{4} \Gamma_q^s D_{qft} + \Gamma log(CCI_{fpt}) + \delta S_{fpt} + \psi_t + \phi_f + \tau_p + \epsilon_{fp,t+1} \tag{5}$$

All the interpretations and explanations of expected signs provided for the parameters in the expression (1) apply entirely for the analogous parameters in the expression (5). The expected sign for Γ is negative, as a lower CCI_{fpt} should reduce the cost of production of p for f in t and this should in turn affect positively the potential profitability.

Very importantly, CCI_{fpt} changes when the prices of inputs change in such a way that it captures in theory the key differential elements (i) and (ii) explained in detail above. As these very granular and specific elements are predicted by the theoretical model presented in the second chapter of this thesis in its more specific and granular proposition, their possible empirical validity when the prices of inputs change exogenously may be interpreted as *causal* evidence in favor of this model.

Estimation of structural parameters

As I mentioned above, it is possible to compute CCI_{fpt}. For this, let us start by writing the first order condition of the cost equation of a firm f to produce a product p in a year t with respect to the material s (this is, with respect to M_{fpst}). As I showed in the expression (17) of the second chapter of this thesis, this first-order condition is as follows[17]:

$$g_p\left(K_{ft}, L_{ft}, H_{ft}\right)^{\beta_p(1-\eta_p)} \frac{\theta_p(1-\eta_p)}{\rho_p} \left[\sum_k \phi_{fkt}\phi_{pk}M_{fpkt}^{\rho_p}\right]^{\frac{\theta_p(1-\eta_p)}{\rho_p}-1} \phi_{fst}\phi_{ps}\rho_p M_{fpst}^{\rho_p-1} = q_{fst} \tag{6}$$

It is possible to identify ϕ_{fst} and ϕ_{ps} from this expression with a combination of some simple transformations and OLS regressions. The expression (6) can be rewritten in terms of the observable revenue perceived by the firm from sales of product p in t (represented by R_{fpt}) as follows:

$$\frac{R_{fpt}}{q_{fst}} = \frac{\alpha_{fpt}}{\theta_p(1-\eta_p)\phi_{fst}\phi_{ps}M_{fpst}^{\rho_p-1}} \tag{7}$$

where $\alpha_{fpt} = \left[\sum_k \phi_{fkt}\phi_{pk}M_{fpkt}^{\rho_p}\right]$. The ratio in the left-hand side of (7) is observable because both the revenues from selling all the products and the prices of the inputs are observables for every firm. I represent such ratio here by y_{fpst}. In addition, I take the natural logarithm of both sides of the equality in (7) to get the following expression:

$$ln y_{fpst} = \psi_{fpt} - \delta_{ps} - \Omega_{fst} - \epsilon_{fpst} \tag{8}$$

where $\psi_{fpt} = log\left[\frac{\alpha_{fpt}}{\theta_p(1-\eta_p)}\right]$, $\delta_{ps} = log\phi_{ps}$, $\Omega_{fst} = log\phi_{fst}$ and $\epsilon_{fpst} = logM_{fpst}^{\rho_p-1}$. The important feature of expression (8) is that all its unknown components can be estimated by running an OLS regression of the observables

[17] Please see the second chapter of this thesis for an explanation of all the parameters and variables involved in the expression (6)

lny_{fpst} as a function of firm-product-year, product-input and firm-input-year fixed effects (which will account for ψ_{fpt}, δ_{ps} and Ω_{fst}, respectively). This is important because by estimating δ_{ps} and Ω_{fst} it is possible to recover two key components of CCI_{fpt}: the firm-input-specific productivities ϕ_{fst} and the input-output coefficients ϕ_{ps}.

I use the OLS estimators of ϕ_{fst} and ϕ_{ps} in this chapter to calculate CCI_{fpt}. In addition, they are used in the section 6 of this paper to analyze possible empirical evidence of the existence of the main mechanism and main assumption of the model presented in the second chapter of this thesis.

Given that I have now estimators for ϕ_{ps} and ϕ_{fst} and the prices paid by all the firms for all the inputs are observables, all I still need to calculate CCI_{fpt} is a value for σ. I take its value from the work from Eslava and Haltiwanger (2020), who also used data from the Colombian EAM and performed a GMM estimation of the elasticity of substitution across materials of Colombian manufacturing firms. They assumed a CES functional form very similar to the one I use here. Because of this comparability in terms of the used data and the functional form, their value is highly applicable to this paper. Their average across all sectors of these elasticities of substitution is 1.84. This is the value that I use here to calculate CCI_{fpt}.

Once calculated, CCI_{fpt} is used in the regression presented in the expression (5) above. Very importantly, CCI_{fpt} can be estimated for the products not produced by the firms, and not only for the products that they actually produce. This is crucial, as I use it in this section in a regression with a dummy for product addition in $t+1$ of products *not produced in* t as the dependent variable. Therefore, it will have to be used by definition for products that are not produced in t.

Possible endogeneity

There may be a problem of endogeneity of the variable CCI_{fpt} in the expression (5). Namely, this variable might be correlated with the variables included in $\epsilon_{fp,t+1}$. For instance, the price that the sellers of inputs charge a firm f for the inputs needed to produce a product p can depend on their own forecast about the future decision by f to produce or not p in $t+1$. If the seller of the inputs forecasts that the firm will not produce p in $t+1$, it might reduce the prices of inputs in t (especially for perishable inputs). If this forecast is based on variables not included in the model (which is likely), this would mean that CCI_{fpt} would be correlated with $\epsilon_{fp,t+1}$, and the OLS estimator of Γ would be biased. To solve this, I perform here a two-stage least squares (2SLS) estimation, using tariffs as instruments.

The Colombian trade reform of 2012

Colombia is a small open economy [18]. This means that it is reasonable

[18] The term "open" means in this context that goods and services can be traded from and to the country, even though some of them are subject to tariffs.

to assume that it is a price taker in the tradable products sold under perfect competition, and that prices of imports should have influence on the prices of the tradable products sold under monopolistic competition. Therefore, if the imports of the tradable inputs used by a firm f become cheaper, the prices of tradable inputs paid by f should fall, even if f buys them from domestic producers. Formally, this means that there should be a positive relationship between the price paid by f for an input k in t (q_{fkt}) and the tariff imposed by the domestic government to input k in t (T_{kt}).

A simple way to justify formally the assumption that q_{fkt} is positively correlated with T_{kt} comes from the optimal pricing behavior of firms under free trade. In general, the price charged by sellers of domestic inputs to manufacturing firms is equal to the marginal cost of the respective input multiplied by a factor F. This factor can depend on several other factors, such as the elasticity of substitution and the number of firms. It is constant under most of the models typically used in economics, and it is equal to one under perfect competition.

Formally, for domestic inputs $q_{fkt} = F.MC_{fkt}$, where MC_{fkt} is the marginal cost of producing the input k faced by the domestic firm that produces it and sells it to f in t. Analogously, the price of imported inputs is $q^*_{fkt} = F.MC^*_{fkt}.(1 + T_{kt})$, where MC^*_{fkt} is the marginal cost of producing the input k faced by the foreign firm that produces it and sells it to f in t. If the tariff T_{kt} decreases, q^*_{fkt} also decreases. Unless the domestic producer of an input k is a monopolist, its price q_{fkt} falls when q^*_{fkt} falls, because F falls for k as a consequence of the higher competition. Under perfect competition, the fall in the tariff is fully transmitted to the price.

Given this, the policy action used here as a source of exogenous variation in the prices of inputs is a unilateral reduction of the tariffs imposed by the Colombian government to the imports from the United States that took place in 2011, prior to the free-trade agreement between these two countries in 2013. As figure 1 shows, the simple average of the product-level tariffs imposed by the Colombian governments to the imports from the United States fell almost 3 percentage points from 2010 to 2011, and kept on falling during almost all the successive years. In eighteen years this average fell almost 8 percentage points, from near 12 percent in 2000 to approximately 4 percent in 2018.

The figure 2 shows that it was not only the average of the product-level tariffs to the imports from United States that changed, but also the distribution of such tariffs. The upper and lower limits of blue boxes in this figure represent the percentiles 25th and 75th of the distribution of tariffs each year, respectively. The horizontal line inside each box is the median of each year. The two horizontal lines above and below each box are the upper adjacent values and the lower adjacent values each year, respectively [19].

Figure 2 shows several facts. The most important is that the distribution

[19] The upper adjacent value is uniquely defined as the j percentile of the distribution such that complies with two requirements: (A) it is smaller than percentile(75)+1.5*(percentile(75)-percentile(25)), and (B) the j+1 percentile is larger than this value. The lower adjacent value is defined in an analogous way

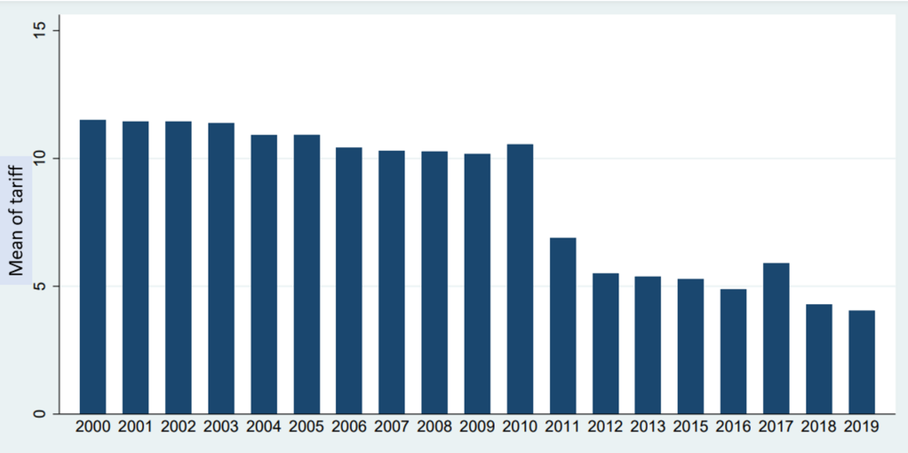


Figure 1. Average of product-level tariffs imposed by the Colombian government to the imports from the United States.

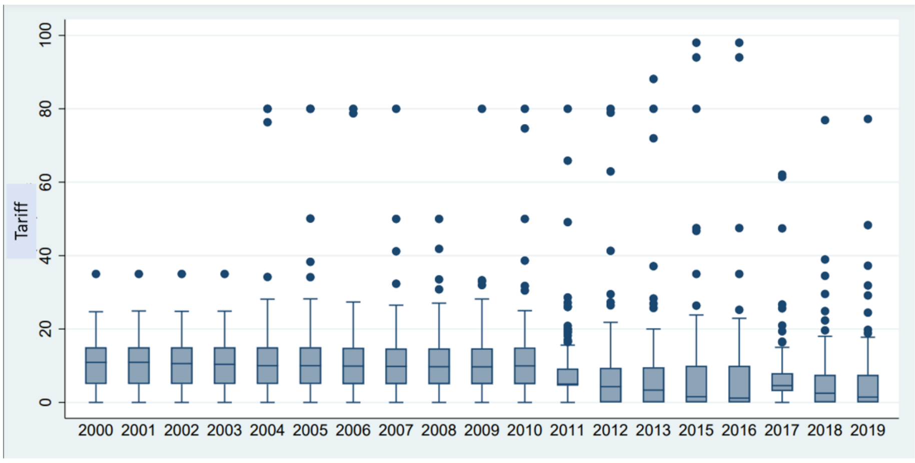


Figure 2. Box plots of yearly distributions of product-level tariffs imposed by the Colombian governments to imports from the United States.

of Colombian tariffs to imports from the United States changed in many senses in 2011. Namely, the interquartile range (75th percentile - 25th percentile) fell from nearly 10 percentage points in all years from 2000 to 2010 to approximately 5 percentage points in 2011. In this last year 50% of the tariffs were between 5 and 10%, which moved the median down from 10% in previous years to 5% in 2011. Several values that would have been below the upper adjacent values in previous years became outliers in 2011. This change in the concentration of the distribution towards lower values continued in the successive years. As a consequence, the median of the distribution fell almost to zero in 2016. Even though it increased temporarily to nearly 3 percent in 2017, it fell again since then until 2019, when it reached again a value very close to zero.

The fact that the median of product-level tariffs imposed by the Colombian government to the imports from the United States fell nearly 5 percentage points from 2010 to 2011 means that many tariffs changed, and that the reduction was not applied just to a few products. This conclusion is reinforced by the other changes in the distribution that can be observed in Figure 2 from 2010 to 2011. The 25th percentile seems to have remained the same (around 5 percent), but the mean fell from nearly 10 percent to nearly 5%. These facts mean that the second quartile of the distribution squeezed. More specifically, 25% of the products had until 2010 a tariff between 5 and 10%, and they (or other products that replaced them in the second quartile) switched to have a tariff of 5% or below.

As a consequence of the squeezing of the lower quartiles described before and of possible reductions of tariffs in the upper parts of the distribution, the 75th percentile also fell nearly 5 percentage points from 2010 to 2011, from nearly 15 percent to 10%. In summary, the change in tariffs that took place in 2011 was not a reduction in just a few products. Instead, it affected a proportion of the products sufficiently high as to cause the notorious change in the distribution that can be seen in Figure 2. The same is true for the years after 2011. Additional checks confirm that the tariffs of more than half of the products fell in 2011, and the same is true for the subsequent years.

The fact that the reduction in tariffs from 2011 was broad both in its extensive (number of affected products) and intensive (extent of changes) margins allows me to use it as a phenomenon that affected most of the products used as inputs by the Colombian manufacturing firms. As for the exogeneity of this phenomenon, it is reasonable to assume that the residuals ϵ_{fpt} in expression (1) in 2011 and in the subsequent years were uncorrelated to the decision of the government to reduce the tariffs of the inputs that are critical for the production of p by firm f since 2011.

Even though some factors that might affect product addition by a firm f such as its capacity of agency or lobbying might be correlated to the decision of the government to reduce or not the tariffs of the inputs used by f, I discard the possibility that this is a generalized fact. This because the median of the market share of a firm in the total market of a product in a year is below 30 percent, even though the definition of product used here is as narrow as possible (that is, I use the most disaggregated level of product definition available in the different product classifications used here).

Results

If the generalized reduction in tariffs that occurred from 2011 led to a fall in the prices of the inputs used by the Colombian manufacturing firms, this generalized reduction should cause a decrease of the right-hand-side of expression (4). Given this relevance as instruments, I use here the average tariffs imposed by the Colombian government to imports from the rest of the world as instruments for CCI_{fpt} in a 2SLS regression.

In the first stage I regress CCI_{fpt} on the two principal components of the tariffs in t on all the inputs used by the firm in t. These two principal components are uncorrelated with each other and they have two key properties: (i) they capture the maximum possible variance of the original tariffs, which means that they capture to a large extent the most notorious differences across tariffs within a firm in a year, and (ii) they are therefore a two-dimensional comprised version (in terms of variance) of the original tariffs. Therefore, they capture to the largest possible extent the variability of these latter while still allowing me to have just two uncorrelated variables derived from *all* the tariffs for each firm.

The two-dimensional nature of principal components allows me to run a unique regression for all the firm-year-product combinations in the first stage. This regression has just two explanatory variables in all cases, unlike the potential situation in which I had used all the relevant tariffs for each firm in each year[20]. In the second stage I used the predicted values from the first stage $C\hat{CI}_{fpt}$ as an explanatory variable in (5).

Table 5 shows the results both using such principal components as instruments for CCI_{fpt} and using CCI_{fpt} itself as a non-instrumented regressor.

I report results in all the specifications included in the table (5) both including and excluding similarity and its interactions with dummies for quartiles of non-production workers. This because there are valid reasons both to include them and to exclude them.

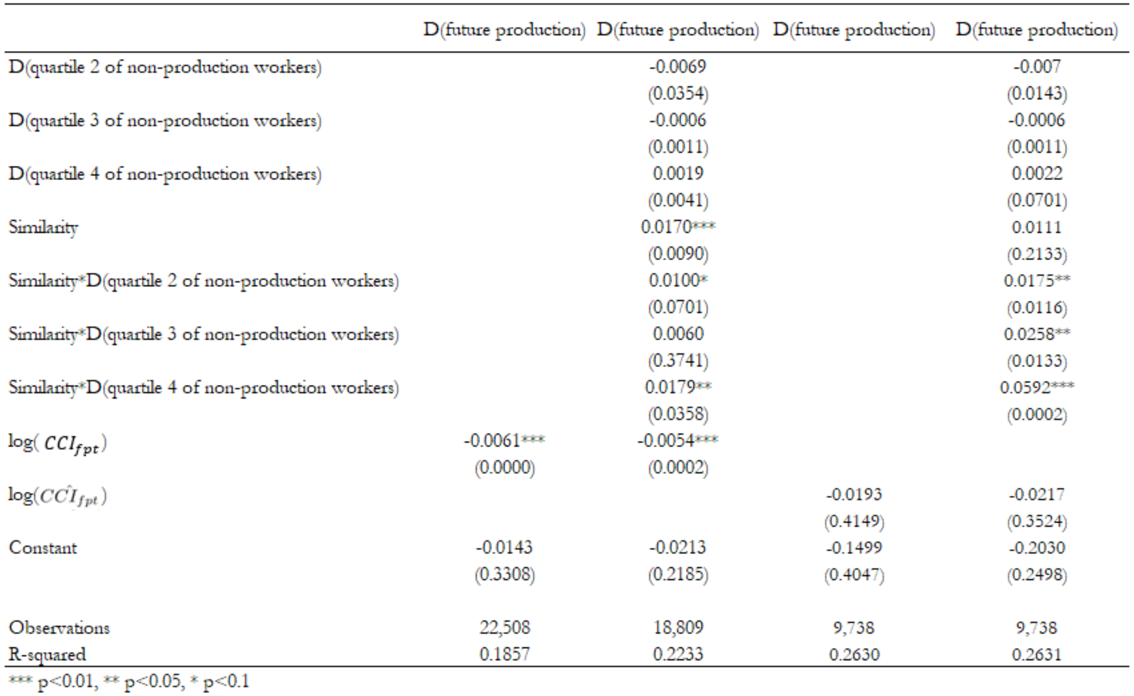
The similarity might be highly correlated with CCI_{fpt}. If a firm f is very proficient in the use of a particular input k in t and the product p is very intensive in k, then CCI_{fpt} will be low. If these assumptions hold, it is likely that f uses k intensively, and also that the production of p is in general intensive in k. In other words, it is likely that the similarity between f and p in the use of inputs is high, which in mathematical terms would mean that S_{fpt} is high. Therefore, there are theoretical reasons to expect a high negative correlation be-

[20] It would be possible in theory to run firm-year-specific regressions of CCI_{fpt} as a function of all the tariffs relevant for the firm in t. However, the number of observations for each firm in this case would be equal to the number of produced actually products by it in t. As each manufacturing firm produces on average less than four products in a year, the number of observations of each regression in this second specification would be very low in many cases. This problem can be reinforced by the fact that some firms use many inputs, which increases the number of parameters to be estimated.

tween CCI_{fpt} and S_{fpt}. The negative correlation between these two variables is indeed slightly above -0.5 in absolute value. Therefore, my estimator for Γ might be inconsistent if I exclude S_{fpt} and its interactions from (5).

However, there are also theoretically valid reasons to exclude S_{fpt} and its interactions from (5), as it is possible to argue that CCI_{fpt} and S_{fpt} are conceptually redundant and they do not need and should not be included in the same regression, as the same fundamentals drive them both. If a firm f is very good at using an input k and some products are very intensive in k, the model presented in the second chapter of this thesis predicts that such products are more likely to be added by f than others because f can produce them efficiently (this is, because CCI_{fpt} is low). This higher relative efficiency *is materialized in reality* by the fact that f uses k intensively, just like k is intensively used to produce those products. This similar intensity implies a high S_{fpt}. In short, similarity is just a metric that *reflects and materializes efficiency*, and its exclusion should not generate any inconsistency, as it does not have influence on the error once CCI_{fpt} is included.

The extent to which the models including similarity and its interactions are better suited to yield a consistent estimator for Γ than those excluding them depends on the extent to which similarity is correlated with other factors related to product addition and uncorrelated to firm-input-specific productivities, such as general technological changes that determine input-output coefficients. As there is not certainty about this extent, I decided to report results both with and without similarity and its interactions in all cases.



	D(future production)	D(future production)	D(future production)	D(future production)
D(quartile 2 of non-production workers)		-0.0069		-0.007
		(0.0354)		(0.0143)
D(quartile 3 of non-production workers)		-0.0006		-0.0006
		(0.0011)		(0.0011)
D(quartile 4 of non-production workers)		0.0019		0.0022
		(0.0041)		(0.0701)
Similarity		0.0170***		0.0111
		(0.0090)		(0.2133)
Similarity*D(quartile 2 of non-production workers)		0.0100*		0.0175**
		(0.0701)		(0.0116)
Similarity*D(quartile 3 of non-production workers)		0.0060		0.0258**
		(0.3741)		(0.0133)
Similarity*D(quartile 4 of non-production workers)		0.0179**		0.0592***
		(0.0358)		(0.0002)
log(CCI_{fpt})	-0.0061***	-0.0054***		
	(0.0000)	(0.0002)		
log($CC\hat{I}_{fpt}$)			-0.0193	-0.0217
			(0.4149)	(0.3524)
Constant	-0.0143	-0.0213	-0.1499	-0.2030
	(0.3308)	(0.2185)	(0.4047)	(0.2498)
Observations	22,508	18,809	9,738	9,738
R-squared	0.1857	0.2233	0.2630	0.2631

*** p<0.01, ** p<0.05, * p<0.1

Table 5. Results of regressions of product addition in $t+1$ as a function of a cost index and similarity in the use of materials interacted with dummies for quartiles of non-production workers. Estimated standard errors one line below the estimators.

The estimator for CCI_{fpt} is negative as expected under all the three specifications, both with and without the inclusion of the similarity and its interactions with the dummies for the quartiles of skilled labour. However, this estimator is statistically significant only when I use OLS without using the principal components of tariffs as instruments for CCI_{fpt} (it is actually highly significant in this case). The number below each estimator is the p-value for the null hypothesis that the corresponding true parameter is equal to zero. The p-values show that the null hypothesis of non-significance is far from being rejected when I use tariffs as instrument for CCI_{fpt}.

If the changes in tariffs are more exogenous than the prices of inputs to the non-modeled firm-product-specific factors that are correlated with product addition (as I reasonably assume), these results suggest that there is not conclusive causal evidence in favor of the model presented in the second chapter of this thesis. However, it is still valid to state that if changes in the prices of inputs were exogenous to such factors to some extent (as might be the case for firms in more competitive markets), the first two columns of Table 5 would indicate the presence of such evidence to some extent.

Very importantly, the results in Table 5 are highly consistent with the prediction derived from the proposition 2 in section 2 that $0 \leq \beta_2^s \leq \beta_3^s \leq \beta_4^s$. More specifically, the results in Table 5 are more consistent with this prediction than those shown in Table 2 in section 4. The only difference is that in Table 5 I included the CES cost index (CCI) as a regressor. As long as this variable is correlated with the product addition and its drivers (structural firm-input-specific productivities and input-output coefficients) are not fully captured by the similarity, its exclusion would generate an omitted variable bias in the results of Table 2 that would not exist in the results of Table 5 (provided that CCI is correlated with the similarity, as is the case). In summary, the results shown in Table 5 validate empirically the proposition 2 in section 2 to a larger extent than the results shown in Table 2, given the correction of a possible omitted variable bias in the former case.

The possibility of using $\hat{CCI}_{fpt}$ to get a consistent estimator of Γ in expression (5) depends on the validity of the tariffs as instruments. Formally, they must be uncorrelated with the errors $\epsilon_{fp,t+1}$. Given that expression (5) includes fixed effects by firm, year and product, $\epsilon_{fp,t+1}$ includes only factors that vary for two or more combinations of those three dimensions. In other words, it includes in theory all the firm-product-specific, firm-year-specific, product-year-specific and firm-product-year-specific factors that have effect on the product addition decisions and that are not included as regressors in expression (5). It is reasonable to assume that all the factors that vary to some extent across firms are uncorrelated with tariffs, as the generalized reduction in tariffs that took place in 2011 was a comprehensive policy that did not attempt primarily to attend requests of specific firms nor to solve firm-specific problems that might affect product addition.

Product-year-specific factors can be more problematic, as there may be time-varying product-specific factors that are common across all the firms that might potentially produce the product in question with two problematic properties:

(i) these factors might affect the firms' decisions to add the product in question, and (ii) these factors might be correlated with the tariffs imposed by the Colombian government to the imports from the U.S. of the inputs needed to produce that product. If (i) and (ii) were true, the tariffs would not be valid instruments.

The possible existence of time-varying product-specific factors with the properties (i) and (ii) above can be rationalized in several ways. Firstly, there might be time-varying product-specific shocks that might be correlated both with the product addition of the product in question and with the tariffs imposed by Colombia to the imports from the U.S. of the inputs needed to produce the product in question, such as demand shocks. If this were the case, the tariffs would not be valid instruments anymore. Secondly, some sectors [21] that produce some specific products might be able to foresee the future changes in the tariffs of the inputs needed for their production, and they might take this into account when deciding to produce their products or not some years before the tariffs reduction. If there is persistence in this "anticipation effect", the tariffs would not be valid instruments. Thirdly, the tariffs reduction might have been more pronounced for those products that are used as inputs of some specific products that are produced by sectors (this is, sets of firms) with specific properties, such as high lobbying capacity. If this capacity is persistent and correlated with product addition, the inputs tariffs would not be valid instruments.

The first possibility is the least worrisome, as I found in the first chapter of this thesis that the change in the average sales price of a product is uncorrelated with the probability that the product in question is added by the firms that may potentially produce it. This should not be the case if demand factors potentially correlated with the tariffs (such as demand shocks) affected the product addition. As for the second and third concerns, I do three things to discard their possible occurrence.

To start, I perform a parallel trends test for the product addition, in order to discard the possibility that the sectors that were more favoured by the inputs tariffs reductions added more products before such reductions than the rest (possibly because they foresaw these reductions). For this test, I use as pre-treatment period the years 2008 and 2009, and as post-treatment period the year 2010. This because the tariffs reduction was carried out in 2011. As for the treatment, I categorize as treated those products for which the weighted average of the tariffs on their inputs (using expenditures shares as weights) are above the median of this metric of inputs tariffs reduction, and as untreated the rest of products (this is, those for which this metric is below its median). The test consists of an augmented difference-in-differences regression. Formally, I ran the following regression:

[21] I use the term "sector" here to make reference to the set of firms that produce a product. I use this term when I need to emphasize the fact that these firms constitute an aggregate with possible specific characteristics such as lobbying capacity. I can use this term and the term "product" indistinctly without incurring into any conceptual mistake, as they both correspond to the same level of aggregation. Please notice that a firm can belong to several sectors here, as firms can be multiproduct.

$$D_{fp,t+1} = \alpha_0 + \alpha_1^{'} X_{fpt} + \alpha_2 D_{pt} + \alpha_3 W_f d_{t0} t + \alpha_4 W_f d_{t1} t + \gamma_p + \gamma_t + \epsilon_{fp,t+1}$$
(9)

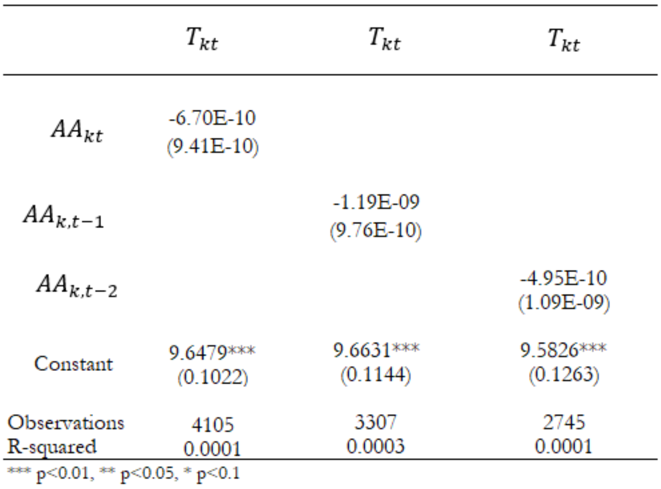
where $D_{fp,t+1}$ is the usual dummy for product addition of product p by firm f one year ahead, X_{fpt} is a set of covariates (the similarity index and its interactions with the quartiles of non-production workers), γ_p and γ_t are product fixed effects and time fixed effects, respectively, D_{pt} is the usual dummy in DID regressions that equals one if the observation corresponds to a treated product (this is, to a product that exhibits a metric of inputs tariffs reduction above its median) and the observation corresponds to 2010 and zero otherwise, w_p equals one if the observation corresponds to a treated product (this is, to a product that exhibits a metric of inputs tariffs reduction above its median), t is a trend, d_{t0} is one for 2008 and 2009 and zero otherwise, and d_{t1} is one for 2010 and zero otherwise.

The term $W_p d_{t0} t$ captures the difference in the linear trend before 2010 (namely, in 2008 and 2009) between the products more favoured by the reduction in the inputs tariffs and the products less favoured by this reduction. Therefore, if α_3 is statistically different from zero, there would be evidence that the trends were different, and it might be the case that sectors that foresaw that would be more favoured by the reduction of the tariffs of their inputs in 2011 started to produce their respective products to a larger extent than the rest of sectors several years before the tariffs reduction. The test consists then in performing a simple Wald test with the null hypothesis that $\alpha_3 = 0$ (parallel trends). I find a p-value of 0.5680 for this test, which allows me conclude that there is evidence of parallel trends in product addition of the sectors more favoured by the reduction of inputs tariffs and those less favoured by such reduction. This solves to some extent the second concern stated above (i.e., the concern of endogeneity because of a possible "anticipation effect").

In order to addess the third concern (possible endogeneity of inputs tariffs reductions because of phenomena such as lobbying capacity that may affect both product addition and tariffs reductions), I adapt the procedure used by Baccini et al. (2019) to the needs of this paper. Namely, I adapt it to take into account that in this paper it is the tariffs on inputs what matter, as these latter are the ones whose principal components I use as instruments. Namely, I run the following regression:

$$T_{kt} = \beta_{0,s} + \beta_{1,s} AA_{k,t-s} + \epsilon_{kt}$$
(10)

where T_{kt} is the tariff imposed by the Colombian government on the imports of input k from the U.S. in year t and $AA_{k,t-s}$ is the addition associated with input k in year $t - s$. It is calculated as follows:



	T_{kt}	T_{kt}	T_{kt}
AA_{kt}	-6.70E-10 (9.41E-10)		
$AA_{k,t-1}$		-1.19E-09 (9.76E-10)	
$AA_{k,t-2}$			-4.95E-10 (1.09E-09)
Constant	9.6479*** (0.1022)	9.6631*** (0.1144)	9.5826*** (0.1263)
Observations	4105	3307	2745
R-squared	0.0001	0.0003	0.0001

*** p<0.01, ** p<0.05, * p<0.1

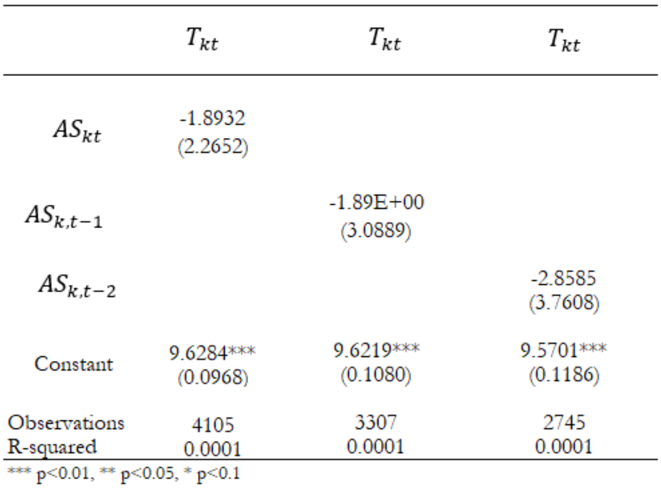
Table 6. Results of regressions of tariffs on the associated addition of products at the input level. Robust standard errors in parentheses.

$$\mathrm{AA}_{k,t-s} = \frac{\sum_{p=1}^{N_k} \Omega_{pk} \sum_f \mathbb{1}_{fp,t-s}}{\sum_{p=1}^{N_k} \Omega_{pk} N_p} \quad (11)$$

where Ω_{pk} is the share of total expenditure of inputs to produce p that is spent on input k, $\mathbb{1}_{fp,t-s}$ is an indicator function that equals one if the potential new product p is added by firm f in $t-s$, N_k is the total number of potential new products that require input k for their production and N_p is the number of firms that have product p in their set of potential new products. Intuitively, $AA_{k,t-s}$ is a number between zero and one that measures the extent to which the products that require input k for their production are added by firms, weighting each of those products by the intensity in which k is used for its production. If I find β_1 to be negative and statistically different from zero, there would be evidence that the inputs that were intensively used in the production of the products chosen by the firms in the past exhibited larger tariffs reductions later, which might be indicative of lobbying capacity or similar phenomena (as long as such capacity manifests in higher product addition to some extent). The table 6 shows the results of the regression above for $s = 0, 1, 2$, in order to explore the relationship between tariffs and product addition for different spans.

I cannot conclude that β_1 is statistically different from zero in any case. This implies that inputs tariffs reductions were not more pronounced for inputs that yielded higher anticipated gains in terms of product addition, which supports the conclusion that non-observables such as lobbying capacity did not seem to determine the changes in tariffs. I run identical regressions with the additional inclusion of an autorregresive component of T_{kt}. The R2 increased to above 0.7. Very importantly, in this case I still cannot conclude that β_1 is statistically different from zero in any case.

Using the product addition as a possible predictor of tariffs might mask the effect of lobbying capacity or similar factors to some extent, as product addition



	T_{kt}	T_{kt}	T_{kt}
AS_{kt}	-1.8932 (2.2652)		
$AS_{k,t-1}$		-1.89E+00 (3.0889)	
$AS_{k,t-2}$			-2.8585 (3.7608)
Constant	9.6284*** (0.0968)	9.6219*** (0.1080)	9.5701*** (0.1186)
Observations	4105	3307	2745
R-squared	0.0001	0.0001	0.0001

*** p<0.01, ** p<0.05, * p<0.1

Table 7. Results of regressions of tariffs on the associated sales at the input level. Robust standard errors in parentheses.

may be influenced by several other factors. In order to surpass this problem to some extent, I repeated the regressions in Table 6 but using the sales instead of the dummies of product addition. Sales have two key properties: (i) I found this variable to be positively correlated with product addition in the first chapter of this thesis (which makes it an observable suspect to cause endogeneity, as I excluded it from the regression in (5)), and (ii) it is reasonable to assume that sectors with higher sales have higher lobbying capacity. With this in mind, I constructed the following variable of associated sales:

$$\text{AS}_{k,t-s} = \textstyle\sum_{p=1}^{N_k} \Omega_{pk} Sales_{p,t-s}(12)$$

where $Sales_{p,t-s}$ is the logarithm of the total sales of product p in year $t-s$. Intuitively, AS measures how large were the sales of the products that require input k for their production, weighting each product by the extent to which k is required to produce it (this is, giving more importance to the products that are more intensive in k). Just as for the case of AA above, I ran regressions of T_{kt} as a function of $AS_{k,t-s}$ for $s = 0, 1, 2$. The results are shown in Table 7. I found the coefficient for AS to be statistically not different from zero in all cases. In words, the sectors with anticipated higher sales did not benefit from larger reductions of the tariffs of their most important inputs. In this case I also repeated the regressions with an autorregresive component of the tariffs. Once again, this did not alter the conclusion of statistical insignificance of the relevant parameter. This finishes my tasks to discard the third concern stated above.

Summarizing, it is reasonable to assume that the principal components of tariffs are valid as instruments for CCI_{fpt}, as it is reasonable to assume that they are uncorrelated with $\epsilon_{fp,t+1}$ in (5).

6 Evidence of the process of learning by using and of the effect of productivities on product addition

In this section I present an additional empirical analysis that explores the possible existence of empirical evidence in favor of the main assumption of the model proposed in the second chapter of this thesis, and also of its main mechanism. To do this I use the structural parameters (input-output coefficients and firm-input-specific productivities) that I identified and estimated in the previous section. The main conclusion from this section will be that the evidence is mostly in favor of the main assumption of the model presented in the second chapter of this thesis, and also of its main mechanism.

Main assumption and main mechanism of the model

The main assumption of the model proposed in the second chapter of this thesis is that there exists a process of learning by using, in which firms increase their firm-input-specific productivities as they use the respective inputs to a larger extent. The main mechanism of this model is that the products that require intensively for their production in t those inputs for whose use a firm f has high firm-input-specific productivities are more likely to be added by f in the years after t. Testing empirically the possible validity of this assumption and mechanism is very important, as I explain below.

What makes special (a) the process of learning by using (main assumption) and (b) the fact that firms add more easily products for which $\sum_k \phi_{pk}\phi_{fkt}$ is larger (main mechanism) is that if (a) and (b) happen in reality, then *necessarily* the finding that the correlation between similarity in t and the product addition after t persists over time is attributable to persistent firm-input-specific productivities to some extent, which is the main feature of the model proposed in the second section of this chapter. This would constitute additional and strong evidence in favor of the model presented in the second chapter of this thesis, in addition to the findings presented in the previous two sections, which are consistent with some of this model's propositions.

Let us assume for explanatory purposes that (a) and (b) occur in reality. This would necessarily imply that the main findings of this work are attributable to persistent firm-input-specific productivities because the interaction of (a) and (b) *necessarily* implies that the firms add more easily more similar products to their product mixes. If a firm f has a high productivity to use k in t, it optimally uses k intensively in t. If (a) happens in reality, then *necessarily* f will have a high productivity to use k in $t+1$. If (b) happens in reality, then f *necessarily* adds to it product mix in $t+1$ more easily those products that require k intensively for their production, and (once again) f uses k intensively to produce them, as it is optimal to do it. Therefore, the products produced by

f in t and the products produced by it in $t+1$ are *necessarily* similar in their input mixes, as they are both intensive in k. Therefore, my empirical finding that the products added by firms in the future have indeed input mixes that are more similar to the input mix used by firms in t is *necessarily* attributable to some extent to the persistent firm-input-specific productivities.

Econometric models

Fortunately, it is possible to run regressions to assess the empirical validity of (a) and (b). As for (b), it is possible to run product-specific regressions of dummies of product addition as a function of $\sum_k \phi_{pk}\phi_{fkt}$ in order to assess the empirical validity of (b). If the estimator for the coefficient for $\sum_k \phi_{pk}\phi_{fkt}$ were positive, I could claim to have found evidence consistent with (b), as this would mean that firms have on average a higher probability of adding products that are intensive in the inputs for whose use they have high productivities. Formally, I run the following regression separately for each p:

$$D_{fp,t+1} = \delta_{0,p} + \delta_{1,p} \sum_k \phi_{pk}\phi_{fkt} + E_{fp,t+1} \tag{13}$$

$D_{fp,t+1}$ equals one if p is added by f in $t+1$, and zero otherwise. This variable is defined on the domain of all the potential new products for f [22]. Parameters ϕ_{pk} and ϕ_{fkt} were estimated in section 5. $E_{fp,t+1}$ is an error term that is assumed to have mean zero and to follow a pattern of variability and correlation across firms and products that is directly estimated from data [23].

If δ_1 is found to be positive for a product p, this would mean that p is added more easily on average by firms with high productivities for the use of the inputs in which the production of p is intensive. If this were the case, this evidence would not be contradictory with (b) above. If this were the case for many products, the evidence would be mostly in favor of (b). I ran a different estimation for each product. This is why both coefficients in (13) are product-specific.

As for (a), I ran input-specific regressions of the firm-input-specific productivity of each firm f in material k in year $t+1$ ($\phi_{fk,t+1}$) as a function of the amount of k used by f in year t ($_{fkt}$). Formally, I ran the following regression:

$$\phi_{fk,t+1} = \theta_{0,k} + \theta_{1,k} M_{fkt} + E_{fk,t+1} \tag{14}$$

$E_{fk,t+1}$ is a firm-material-year-specific error term that is assumed to have the same properties as $E_{fp,t+1}$ above. A positive estimator of θ_1 would mean that it does happen on average for input k that the productivities of firms to

[22] See section 2 for a definition of the potential new products

[23] In short, this means that I use the Huber-White sandwich estimator of the variance-covariance matrix

use it grow on average as they use it more (by the process of learning by using). This is exactly what (a) states. If this were the case for many inputs, I would have found evidence that is generally consistent with (a). I ran a different estimation for each input. This is why both coefficients in (14) are input-specific.

Distribution of the productivities and of the input-output coefficients

There are two key variables involved in the expressions (13) and (14) that are not directly observable from data: ϕ_{pk} and ϕ_{fkt}. I found estimators for them in section (5). Figures 7 and 8 show the distributions of the estimators of ϕ_{pk} and ϕ_{fkt}, respectively.

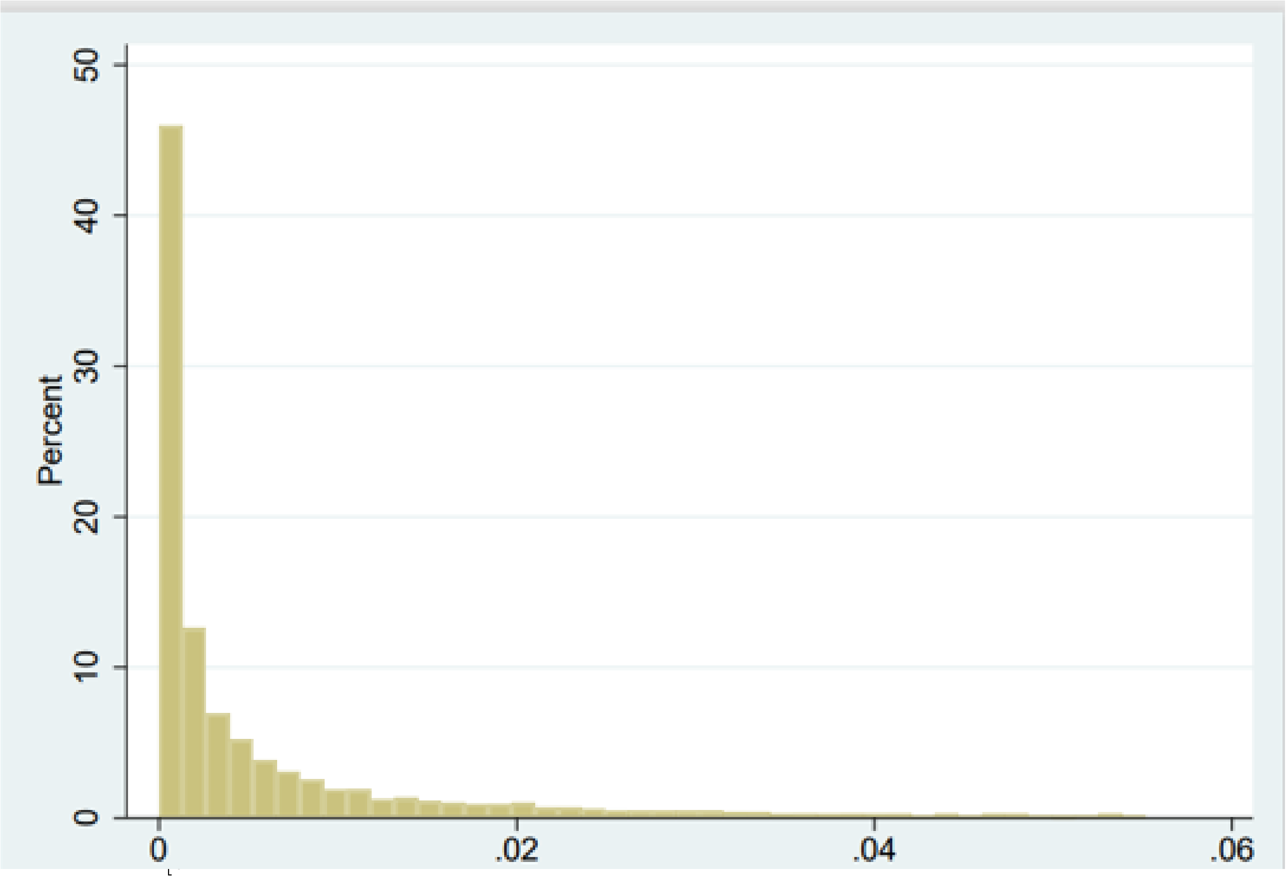


Figure 3. Histogram of estimators of input-output coefficients ϕ_{pk}. It shows the percentage of each interval.

The estimators of ϕ_{pk} and ϕ_{fkt} shown in Figures 3 and 4 are both concentrated in low values. The distribution is in both cases highly skewed towards the lowest intervals. There are also some firms and products with intermediate values of firm-input-specific productivities and input-output coefficients, respectively. Finally, there are a few firms and products with high firm-input-specific

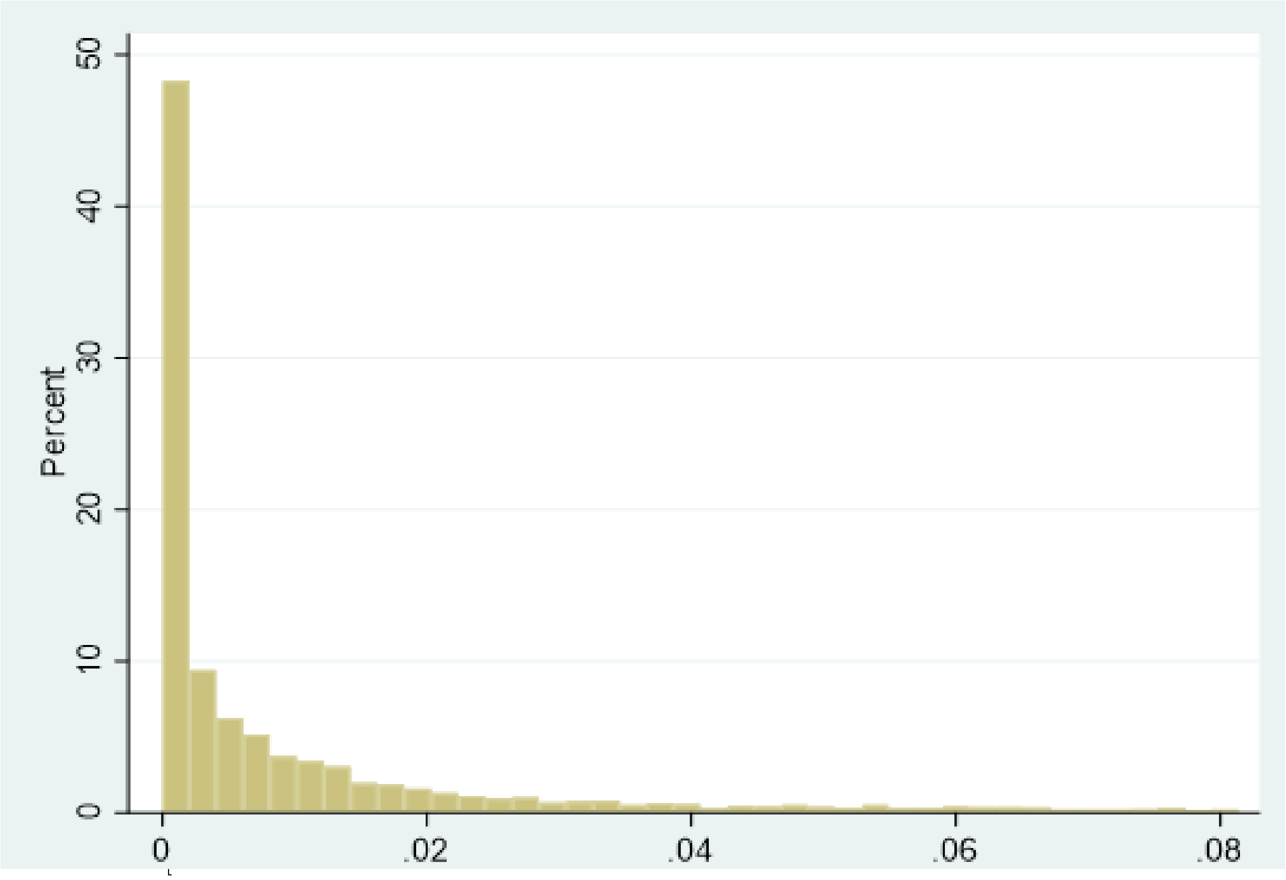


Figure 4. Histogram of firm-input-specific productivities ϕ_{fkt}. It shows the percentage of each interval.

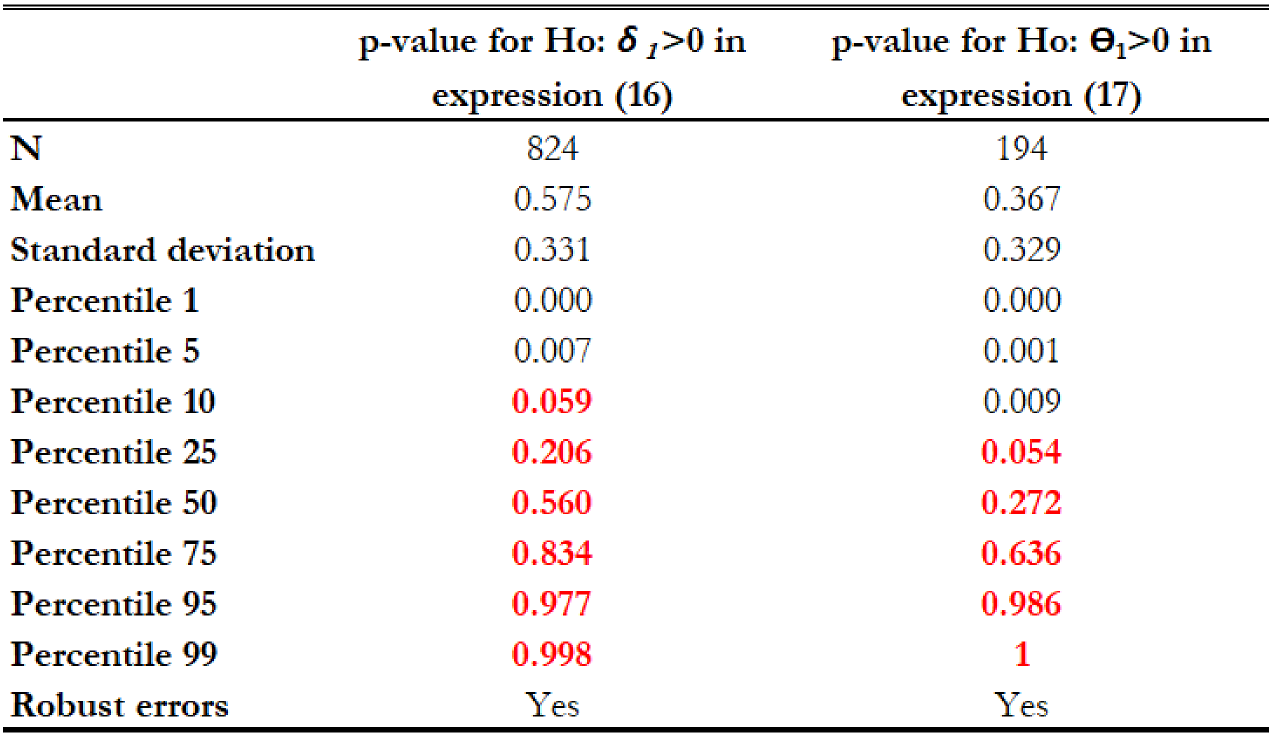
productivities and input-output coefficients. In summary, the contributions of inputs to production and the firm-input-specific productivities of firms are both small in most cases, and there are just a few cases in which inputs contribute to a large extent to production or firms are very productive at using inputs. This is qualitatively consistent with the finding from Del Gatto et al. (2006), who found that productivity has an empirical distribution that can be characterized as a theoretical Pareto distribution, as this latter is also skewed towards low values, and exhibits very high values at its upper tail.

Results

I ran all the feasible product-specific and input-specific regressions of expressions (13) and (14), respectively. By feasible I mean that I ran all those regressions for which I had a sufficient number of observations. This restriction allowed me to run 824 product-specific regressions and 194 input-specific regressions, as shown in Figure 5. As the estimators can be not compared across products and inputs, they are not shown here. Instead, I present in Figure

5 summary statistics that characterize in each case the distribution of the p-values for the null hypothesis that the respective true parameter is positive. More specifically, left column of Figure 5 shows the percentiles and other statistics of the 824 p-values of the 824 test statistics for the null hypothesis that δ_1, p is positive for the 824 products for which I ran regressions for expression (13). Analogously, right column of Figure 6 shows the percentiles and other statistics of the 194 p-values of the 194 test statistics for the hypothesis that θ_1, k is positive for the 194 products for which I ran regressions for expression (14).

Positive values of true parameters δ_1, p and θ_1, k would constitute evidence in favor of (a) and (b). The main conclusion from Figure 5 is that this is mostly the case. More specifically, left column shows that for more than 90 percent of products the evidence does not allow a rejection of the hypothesis that the probability of addition is higher for the products that are intensive in the inputs for whose use firms have higher productivities ((b) above). Similarly, right column indicates that for more than 75 percent of inputs the evidence does not allow a rejection of the hypothesis that the firm-input-specific productivities grow as the respective inputs are used to a larger extent. In summary, the evidence is mostly in favor of (a) and (b).



	p-value for Ho: δ_1>0 in expression (16)	p-value for Ho: Θ_1>0 in expression (17)
N	824	194
Mean	0.575	0.367
Standard deviation	0.331	0.329
Percentile 1	0.000	0.000
Percentile 5	0.007	0.001
Percentile 10	**0.059**	0.009
Percentile 25	**0.206**	**0.054**
Percentile 50	**0.560**	**0.272**
Percentile 75	**0.834**	**0.636**
Percentile 95	**0.977**	**0.986**
Percentile 99	**0.998**	**1**
Robust errors	Yes	Yes

Figure 5. p-values of product-specific and input-specific estimators of slopes of expressions (9) and (10)

7 Conclusions

This paper uses a random sample of a comprehensive dataset of Colombian manufacturing firms to analyze empirically the phenomenon of product addition and its determinants at the firm level. Its main conclusion is that the analyzed empirical evidence is mostly in favor of the most important predictions of the theoretical model of product addition and persistent firm-input-specific productivities proposed in the second chapter of this thesis.

To start, the main regressions of this paper yield evidence in favor of the first proposition of the model presented in the second chapter of this thesis. Namely, I found that firms add on average more easily in the future new products that are more similar to their current production in terms of the input mix, and that this correlation is stronger in the firms with the highest levels of skilled labour (proxied by the number of non-production workers). Very importantly, I found stronger empirical support for the effect of skilled labour on the correlation between similarity and product addition after correcting a possible omitted variable bias caused by the exclusion of the potential cost. I also found evidence that the correlation between current similarity and future product addition remains over time. In other words, the current similarity in the use of inputs is positively correlated with product addition in the future, even 5 years ahead for the firms with the highest values of non-production workers.

I interpret these results as primary evidence in favor of the key role of the interaction of firm-input-specific productivities and input-output coefficients as in the model I proposed in the second chapter of this thesis. More specifically, in that model, firms accumulate over time persistent productivities to use specific inputs as they use these inputs in their productive processes, and this allows them to add more easily in every year products that use those inputs intensively.

In order to establish if the findings explained so far can indeed be attributed to the persistence of firm-input-specific productivities, I do two additional things in this work. Firstly, I used a source of exogenous variation in the prices of inputs to analyze possible causal evidence in favor of the model proposed in the second chapter of this thesis, which relies on these persistent firm-input-specific productivities. Secondly, I used structurally estimated firm-input-specific productivities and input-output coefficients to test empirically the validity of the most important assumption of the model proposed in the second chapter of this thesis, and also of its main mechanism.

For the analysis of causality I use a generalized unilateral reduction of tariffs to the imports from the U.S. carried out by the Colombian government in 2011. This reduction was deepened in the subsequent years. The model proposed in the second chapter of this thesis predicts that this general reduction of the prices of inputs should have affected the product addition of different potential new products by different firms to different extents, given the pre-change firm-input-specific productivities and the input-output coefficients for each firm-product combination. The granularity of this prediction allows me to test for possible causal evidence in favor of the model proposed in the second chapter of this thesis by running a regression for a firm-product-year dummy for product addition that incorporates a structural variable that accounts for the firm-product-specific effect mentioned above. The estimator for this variable is negative but it is not statistically significant in the specifications that correct for possible endogeneity. I conclude from this that there does not exist conclusive causal evidence in favor of the model proposed in the second chapter of this thesis.

As for the direct testing of the main mechanism and the main assumption of the model presented in the second chapter of this thesis, I use structurally estimated parameters to run regressions whose coefficients capture such mechanism and assumption. The results are mostly in favor of the existence of these latter.

In summary, I find in this paper primary evidence in favor of the explanation provided by the second chapter of this thesis for the phenomenon of product addition by firms. Such explanation states that firms add more easily products that are intensive in the inputs for whose use they have higher firm-input-specific productivities. As these productivities remain over time to some extent, they lead firms to produce products that are similar across years in terms of their input mixes. In order to establish additional and more solid evidence in favor of this explanation, I carry out a causality analysis and a structural estimation of key parameters of the model, which I subsequently use also to test directly the empirical validity of the key mechanism behind such explanation, and of the main mechanism of the model. This latter task yields evidence in favor of the model presented in the second chapter of this thesis, whereas the causality analysis does not yield conclusive results.

The results summarized here can be used to inform the process of design and implementation of growth policies. This because these policies should aim at boosting the process of product addition at the firm level, as this would result in gains in terms of efficiency and growth. If the agencies in charge of promoting growth decided to effectively boost the process of product addition at the firm level, they would face the key and non-trivial question of how to do this. The results found in this paper contribute to answering this question to some extent.

The main findings from this paper can be summarized as follows in terms of the dimensions that matter for public policies and that may be affected by these latter: (a) There is path dependence in the process of product addition by firms, as they move more easily to more similar products in terms of the input mix. (b) the capacities of firms to use different inputs determine both the paths chosen by them to expand their product mixes and the speed at which they go through these paths. (c) Having more skilled labour increases the speed at which a firm goes through its path, whatever this latter is. (d) Paths are reinforcing: if a firm chooses a path of products that are intensive in an input, this decision makes this firm even more prone to keep going through this path. (e) Given (d), reaching paths that contain less similar products may take firms longer, unless their capacities to use the inputs in which these products are intensive increase because of a reason that is external to the firm (such as a public policy).

Given this context, there are two main courses of action that national and sub-national authorities in charge of designing and implementing growth policies can take. Firstly, they can implement policies that increase the speed at which firms go through the paths they choose. Secondly, they can help firms to reach new paths of diversification (through product addition).

In order to increase the speed at which firms go through the paths they optimally choose, the authorities can do three things. Firstly, they can facilitate the hiring of the relevant skilled labour by firms. I found in this thesis that having more skilled labour in general increases the speed at which firms go through their optimal paths, but it remains a pending task for each authority to establish what types of skilled labour are critical in each context. Once this is done, the authorities can implement programs that subsidize the hiring of these critical skilled workers conditional on them working on activities that are closely related to product addition. Secondly, the authorities can implement policies that help firms to maintain their existing capacities to use inputs to larger extents. This can be done by improving learning protocols and standardization, facilitating the spreading of input-specific knowledge within and across firms and mitigating the losses of knowledge caused by the turnover of production workers. Thirdly, the authorities can focus their training programmes on identifying the input-specific capacities of each firm, and then on increasing these capacities. Product innovation programmes can focus on activities that aim at implementing new ways of using the inputs identified as critical for each firm.

On the other hand, authorities might be interested in helping firms to reach new paths of diversification. This might be the case if the authorities identify key new industries that they want the country to develop for different economic or non-economic reasons, but that cannot be developed in the desired time horizon with the current state of capacities of firms to use their inputs. In short, authorities might be interested in helping firms to transit from their current paths to other "better" paths (whatever "better" means in each context) in an established time horizon. This thesis suggests possible courses of action to get this. Authorities can implement policies to increase exogenously the input-specific capacities of firms to use inputs that they do not use as frequently and proficiently as others. This can be done by prioritizing the access of firms to these inputs and their productive experimentation with them in the training programmes and also in the knowledge and technology transfer programmes. Partial subsidies to the use of the inputs identified as critical for the key new industries can be also implemented. This policy should yield tangible results, as I found that the costs of production are indeed correlated with product addition, as expected. This thesis does not imply that policies that aim directly at promoting the development of new products (without focusing on the use of the inputs that are critical for their production) are incorrect, but only that focusing on increasing the firms' capacities to use critical inputs can effectively help firms to produce those new products.

Path dependence and the intuition that skills determine what paths are chosen by countries and firms and how easily these go through their chosen paths are not exclusive features of this work. Hausman and Hidalgo (2008), Feenstra and Rose (2000) and several other authors arrived to very similar conclusions and conjectures (please see the first chapter of this thesis for a more detailed discussion about this topic). However, this work offers more specific findings by analyzing more granular phenomena, and this allows me to yield more specific recommendations. Namely, my findings about the existence of persistent firm-input-specific productivities, their dependence on the actual use of the respective inputs, the effect of skilled labour on their persistence and their effect on product addition allows me suggest more specific courses of action, as I did in the previous paragraphs. Future work might focus on analyzing more precisely what types of skilled labour are critical in each context and what policies allow firms to maintain and boost their firm-input-specific productivities to larger extents in each context.

References

Angrist, Joshua D., and Pischke, J. "Mostly Harmless Econometrics: An Empiricist's Companion", *Princeton: Princeton University Press.*

Baccini, Leonardo, Impullitti, Giammario and Malesky, Edmund J. (2019). "Globalization and state capitalism: Assessing Vietnam's accession to the WTO", *Journal of International Economics*, 119: 75-92.

Boehm, Johannes, Dhingra, S. and Morrow, J. (2019). "The Comparative Advantage of Firms", *CEPR discussion papers*, 13699.

Cameron, Colin, Gelbach, J. and Miller, D. (2008). "Bootstrap-Based Improvements for Inference with Clustered Error", *The Review of Economics and Statistics*, 90 (3): 414-427.

Correia, Sergio (2016). "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator", *Working Paper.*

Del Gatto, Massimo, Mion, G. and Ottaviano, G. (2006), "Trade Integration, Firm Selection and the Costs of Non-Europe", *Mimeo, University of Bologna.*

Ding, Xiang. (2020), "Industry Linkages from Joint Production", *Unpublished.*

García García, J., Montes, E. and Giraldo, I. (editors) (2019), "Comercio Exterior en Colombia. Política, Instituciones, Costos y Resultados", *CEP, Banco de la República de Colombia.*

Klette, Tor Jakob (1994), "R and D, Scope Economies, and Plant Performance", *The RAND Journal of Economics*, Vol. 27, No. 3, 502-522.

MacDonald, J. (1985), "R and D and the Directions of Diversification", *The Review of Economics and Statistics*, Vol. 67.

Melitz, Marc, and Ottaviano, G. (2008). "Market size, Trade, and Productivity", *Review of Economic Studies*, Vol. 75, 295-316.

Sutton, John (2012). "Competing in capabilities. The Globalization Process", *Oxford University Press.*

References

Amiti, Mary and Konings, Jozef (2007). "Trade Liberalization, Intermediate Inputs, and Productivity: Evidence from Indonesia", American Economic Review, Vol. 97, No. 5.

Angrist, Joshua D., and Pischke, J. "Mostly Harmless Econometrics: An Empiricist's Companion", Princeton: Princeton University Press.

Baccini, Leonardo, Impullitti, Giammario and Malesky, Edmund J. (2019). "Globalization and state capitalism: Assessing Vietnam's accession to

the WTO", \textit{Journal of International Economics}, 119: 75-92.\\

Bernard, Andrew, Redding, S. and Schott, P. (2007). "Comparative Advantage and Heterogeneous Firms", Review of Economic Studies, 74, 31-66.

Bernard, Andrew and Redding, Stephen (2010). "Multiple-Product Firms and Product Switching", American Economic Review, Vol. 100, No. 1.

Bernard, Andrew, Redding, S. and Schott, P. (2011). "Multiproduct Firms and Trade Liberalization", Quarterly Journal of Economics, Vol. 126, No. 3, 1271-1318.

Boehm, Johannes, Dhingra, S. and Morrow, J. (2019). "The Comparative Advantage of Firms", CEPR discussion papers, 13699.

Broda, Christian and Weinstein, D. E. (2010). "Product creation and destruction: Evidence andprice implications", American Economic Review, 100(3):691–723.

Cameron, Colin, Gelbach, J. and Miller, D. (2008). "Bootstrap-Based Improvements for Inference with Clustered Error", The Review of Economics and Statistics, 90 (3): 414-427.

Cohen, Wesley M. and Klepper, S. (1996), "Firm Size and the Nature of Innovation within Industries: The Case of Process and Product R&D", The Review of Economics and Statistics, Vol. 78, No. 2, 232-243.

Correia, Sergio (2016). "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator", Working Paper.

Del Gatto, Massimo, Mion, G. and Ottaviano, G. (2006), "Trade Integration, Firm Selection and the Costs of Non-Europe", Mimeo, University of Bologna.

Ding, Xiang. (2020), "Industry Linkages from Joint Production", Unpublished.

Eaton, Jonathan, Eslava, M., Kugler, M. and Tybout, J. (2011). "Export Dynamics in Colombia: Firm-Level Evidence", NBER Working Papers, 13531.

Feenstra, Robert C. and Ma, H (2007). "Optimal Choice of Product Scope for Multiproduct Firms under Monopolistic Competition", NBER Working Papers, 13703.

Feenstra, Robert C. and Rose, A (2000). "Putting Things In Order: Trade Dynamics And Product Cycles", The Review of Economics and Statistics, MIT Press, Vol. 82(3), 369-382.

Freund, Caroline and Pierola, Martha (2016). "The Origind and Dynamics of Export Superstars", IDB Working Papers Series, IDB-WP-741.

García García, J., Montes, E. and Giraldo, I. (editors) (2019), "Comercio Exterior en Colombia. Política, Instituciones, Costos y Resultados", CEP, Banco de la República de Colombia.

Guo, Diyue. (2019), "Multiproduct Firms and the Business Cycle", Working Papers Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, 2019-05-01.

Hausmann, Ricardo and Hidalgo, C. (2009). "Country Diversification, Product Ubiquity, and Economic Divergence", CID Working Paper, Harvard University, RWP10-045.

Huergo, Elena and Jaumandreu, Jordi (2004). "How Does Probability of Innovation Change with Firm Age?", Small Business Economics, 22, 193-207.

Jovanovic, Boyan (1993). "The Diversification of Production", Brookings Papers on Economic Activity, vol. 24(1 Microec), 197-247.

Kamien, Morton I. and Schwartz, N (1975). "Market Structure and Innovation: A Survey", Journal of Economic Literature, Vol. 13, No. 1, 1-37.

Klepper, Steven (1996). "Entry, Exit, Growth, and Innovation over the Product Life Cycle", American Economic Review, Vol. 86, No. 3, 562-583.

Klette, Tor Jakob (1994), "R&D, Scope Economies, and Plant Performance", The RAND Journal of Economics, Vol. 27, No. 3, 502-522.

Levinthal, Daniel A. and March, James G. (1993). "The Myopia of Learning", \textit{Strategic Management Journal}, vol 14, 95-112.\\

MacDonald, J. (1985), "R&D and the Directions of Diversification", The Review of Economics and Statistics, Vol. 67.

Melitz, Marc, and Ottaviano, G. (2008). "Market size, Trade, and Productivity", Review of Economic Studies, Vol. 75, 295-316.

Nadiri, M. Ishaq (1993). "Innovations and Technological Spillovers", NBER Working Papers, 4423.

Ornaghi, Carmine (2006). "Spillovers in Product and Process Innovation: Evidence from Manufacturing Firms", International Journal of Industrial Organization, Vol. 24, No. 2, 349-380.

Simon, Herbert A. (1997). "Models of Bounded Rationality. Empirically Grounded Economic Reason", \textit{The MIT press}.\\

Stokey, Nancy L. (1988). "Learning by Doing and the Introduction of New Goods", Journal of Political Economy, Vol. 96, No. 4, 701-717.

Sun, Xiuli, Haizheng Li and Vivek Ghosal (2020). "Firm-level human capital and innovation: Evidence from China", China Economic Review, Vol. 59.

Sutton, John (2012). "Competing in capabilities. The Globalization Process", Oxford University Press.

Vernon, Raymond (1966). "International Investment and International Trade in the Product Cycle", The Quarterly Journal of Economics, Vol. 80, No. 2, 190-207.