ESCUELA DE NEGOCIOS Y ECONOMÍA

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Working Papers

2017-01

‘Distance from the core’ and new export survival: Evidence from Chilean exporters

Daniel Goya

Pontificia Universidad Católica de Valparaíso

Andrés Zahler

Universidad Diego Portales

ene.pucv.cl/wp

'Distance from the core' and new export survival: Evidence from Chilean exporters*

Daniel Goya† and Andrés Zahler‡

October 30, 2017

Abstract

An important part of aggregate export growth is due to firms adding new varieties to their export baskets. We show that a measure of the 'distance' between a new export and the previous export basket is a significant determinant of the survival of a new firm-product export flow. We present evidence suggesting that the measure we use is a good proxy for the theoretical 'distance from the core' in Eckel and Neary (2010) and that this measure captures technological rather than demand complementarities across products.

Keywords: Export diversification; extensive margin; export survival; product proximity.

JEL codes: L25, F14, O30.

1 Introduction

The availability of detailed micro-data has allowed research on the different margins of trade to evolve from the initial country-level studies (e.g., Evenett and Venables, 2002; Hummels and Klenow, 2005) to studies that analyse how firms export different varieties to different destination markets (e.g., Bernard et al., 2009; Gorg et al., 2012; Fontagné et al., 2016, among others). This evolution has allowed us to understand the relevance and workings of these different margins.

As shown by Bernard et al. (2010), almost a third of the 65% growth in US manufacturing output between 1972 and 1997 can be attributed to existing firms adding and dropping products. Navarro (2012) shows that for Chilean manufacturers, this margin represents 44% of the country's manufacturing growth for 1996-2003. Approximately 48% of the 500% growth in Chilean non-mining exports during 1992-2006 is explained by products added by firms that were already exporters.[1]

*The authors would like to thank Beata Javorcik, Meredith Crowley, Giancarlo Corsetti, Peter Neary, Giammario Impullitti, Maurizio Zanardi, Matías Caamaño, Lu Han and Andrei Potlogea for helpful comments and suggestions. We also thank the participants at PhD workshops in Cambridge, the 7th MEIDE Conference in Santiago, the ETSG 2015 Conference, the SECHI 2017 meeting and seminar participants at Universidad de Chile. Both authors acknowledge partial funding from the Nucleo Milenio Initiative NS100017 'Intelis Centre'. Goya also thanks funding from CONICYT (Becas Chile 79090016).

†School of Business and Economics, Pontificia Universidad Católica de Valparaíso, daniel.goya@pucv.cl.

‡Institute for Public Policy, Diego Portales University.

[1]This figure is calculated from our customs data, which will be described in detail later.

This paper examines firms that add new products to their export baskets and studies their survival. In particular, this work analyses whether the likelihood of success of a new export product depends on how 'distant' it is to the products that a firm was already exporting. Our question is motivated by the idea that similar products require similar capabilities, and thus, successfully producing 'distant' products implies higher challenges for the firm. This is related to the concept of 'core competence' used in the multi-product firm model by Eckel and Neary, 2010, and our results are consistent with their model's predictions.

More specifically, we find that the survival of a new product not only depends on firm- and product-level characteristics but also on the *distance* (difference in implicit capabilities) between the new product and the firm's previous export basket, which is a firm-product characteristic.

Following the spirit of Prahalad and Hamel (1990)–who introduced the idea of 'core competences' in the management literature–we expand upon Eckel and Neary's logic, arguing that firms can have multiple 'core competences' or capabilities[2] and that specific combinations are required to produce and export particular products efficiently. We argue that a firm's export basket reveals information about the firm's competences, and we produce evidence supporting this idea.

We build a measure of *distance* using the measure of *similarity* between product pairs defined by Hidalgo et al. (2007). The measure we propose can be interpreted as a proxy for the theoretical 'distance from the core', which takes into account the fact that different underlying capabilities are required for manufacturing different goods.

We use firm-level customs data from Chile for 1991-2006, and we show that the distance between a firm's certain new export product and its *previous year's export basket* is a significant determinant of whether the new exported product is likely to survive beyond its first year of export. We use linear probability models with fixed effects and duration analysis models with random effects, and we find that, *ceteris paribus,* a new export that is at a larger distance from the firm's previous exports is less likely to survive than a new export that is *closer* to the original export basket. More specifically, we find that a 'jump' that is one standard deviation longer than another has a 12% lower probability of surviving more than a year in export markets. This finding is robust to firm-level characteristics, product-level characteristics and even firm-time and product-time fixed effects.

Hidalgo et al. (2007) show that countries change their specialisation patterns gradually, moving toward products that are at a short distance from their current specialisation. This paper adds a micro, within-firm element, showing that firms are *also* more likely to be successful when they move gradually and suggesting that the distance measure proposed by Hidalgo et al. (2007) has useful information about firm-level productive capabilities. Understanding the relationship between new product export decisions and their associated survival dynamics contributes to our comprehension of the micro-level processes that underlie changes in the export basket at the country level.

To reinforce our interpretation of the results, we build measures of distance in inputs and in buyers using manufacturing census data for Chile and show that only input similarity matters.

[2]The concept of capabilities is similar and is more commonly used in the evolutionary economics literature stemming from Nelson and Winter (1982).

This result provides evidence that the measures we use are indicative of the differential *production* capabilities or competences required to produce and export different goods within a firm and suggests the existence of technological rather than demand complementarities. Additional results with domestic manufacturing also support this view.

The paper is structured as follows: Section 2 reviews related theoretical and empirical literature. Section 3 describes the data we use and discusses the definition of the distance variable, showing how the measure proposed is consistent with theoretical models. Section 4 presents the main result, that the risk of failure is higher for new exports that are more distant from the firm's previous export basket. Section 4 also presents a number of robustness tests. Finally, Section 5 summarises the paper's findings.

2 Previous Research

Theory

There have been several approaches to modelling the production and export behaviour of multi-product firms, two of which are particularly relevant for our question. Bernard, Redding, and Schott (2011) developed a model in which productivity in a given variety is determined by two parameters: a firm-level measure of *'ability'* and a firm-product specific *'expertise'*, which is uncorrelated across products.[3] Eckel and Neary (2010) used a different approach that assumes that each firm has a *'core competence'*. In this model, there is a continuum of product varieties. A given core competence locates a firm on that continuum. Products at a larger distance from the firm's core competence are produced less efficiently (with a higher marginal cost).[4] The products at a higher risk of being abandoned after a negative shock to a firm are those that are further away from its core (Eckel and Neary, 2010) or those with the lowest product-specific expertise (Bernard et al., 2011).

To understand whether survival depends on relationships between products, each model contributes one important point but excludes another. Bernard et al.'s (2011) model is interesting in that it implicitly recognises that firms require different capabilities to produce different goods. However, expertise is orthogonal across goods, leaving no room for heterogeneous technological relationships between products due to, for example, common expertise and ability or the capabilities required to produce and export a certain pair of goods. Eckel and Neary's (2010) model defines a distance between different varieties. However, since all of the varieties are placed along a single continuum, the possibility of multiple underlying capabilities is ruled out. Prahalad and Hamel (1990) refer to 'core competences', in plural, and argue that firms use different combinations of core competences for different goods. A nice example that they provide is Canon: this firm's core competences are precision mechanics, fine optics, and microelectronics. The authors argue that '*every Canon product is the result of at least one core competency*'.[5] Our work builds on these ideas, provides evidence of multiple capabilities related to different export products

[3] These can also be interpreted as firm productivity or consumer taste parameters.

[4] Mayer et al. (2014) develop a model that employs a similar idea of 'core competence' but uses monopolistic instead of oligopolistic competition.

[5] The authors use the terms competence and competency interchangeably.

and relates the survival of such products to how far away they are from *core* products.

Empirical Literature - Survival

We review the empirical literature on export survival to inform our empirical specifications. The initial contributions to this literature were made by Besedeš and Prusa (2006a, 2006b). These works find that US country-product import flows tend to be short lived, but those that survive have long tenures. The authors test the matching model of trade formation by Rauch and Watson (2003), confirm its prediction that duration increases with the value exported the first year, and find that homogeneous goods have higher hazard rates.

Brenton et al. (2009) examine the survival rate of new export relationships (defined as product-destination pairs) across 82 exporting countries and find that the country having previous experience with a product and a market, as well as the initial value of a trade flow, are important determinants of survival. Iacovone and Javorcik (2008) show that Mexican manufacturers are more likely to drop export products that have shorter tenure, lower sales, or represent a smaller share of the firm's exports. Gorg et al. (2012) explore the determinants of survival of firm-product export flows from Hungary and find that the probability of survival increases with firm-level total factor productivity, the initial share of the product in the export basket, and tenure and that this probability decreases with the relative unit value of the product. Lejour (2013) conducts a similar exercise with Dutch data and confirms the importance of the initial export value. Córcoles et al. (2014) look at value chains for the automobile industry and find that more 'sophisticated' products have a lower risk of failure. Gorg et al. (2012) consider indicators that proxy for whether a product belongs to the firm's *core* considering the relative importance of the new product in the export basket, but the authors do not define a measure of distance between products.

Two related papers also work with Chilean data. Using firm-level manufacturing census data, Navarro (2012) finds that the probability of stopping the production of a good depends negatively on its value and tenure and positively on firm size, age and the number of products manufactured. Fernandes and Paunov (2015) also look at Chilean manufacturers and find that firms that add new products have higher survival odds as long as they have diversified sources of revenue. Interestingly, the authors find that this relationship is independent of the 'distance' between the new products and their previous baskets—which are defined in a very similar in this paper. This is one of the few studies that consider a measure of proximity between the products exported, but their work looks at *firm* survival. To the best of the authors' knowledge, measures of proximity between the new exports and the previous export basket have not been included in the studies looking at the survival of new *firm-product* export flows.

Not focusing on survival but closely related to our paper, Fontagné et al. (2016) find evidence consistent with the existence of technological or demand complementarities across the products exported by Italian and French firms.

Empirical Literature - Distance

Methodologically, there have been a number of attempts to define measures of similarity

between products. Three approaches can be highlighted. The first is simply the use of classification codes, such as those in the Standard Industrial Classification (SIC), the Standard International Trade Classification (SITC) or the Harmonised System (HS). These approaches assume that similar products are those under the same broad 2- or 4-digit categories. The more digits that two products have in common, the more related they are. A second approach is based on exploiting additional information, for example, the input-output profiles (Fan and Lang, 2000), the type of labour used by firms (Farjoun, 1994), or the classification of patents (e.g., Engelsman and van Raan, 1994). Finally, there are measures of similarity based on the co-occurrence of products or exports of particular products within a certain unit (country or firm). This is the approach used by Teece et al. (1994), Hidalgo et al. (2007) and Neffke and Svensson Henning (2008).

Measures based on a classification system can be as detailed as the classification, but they require the strong assumption that the classification is related to some underlying similarity between products. These measures are also limited in that they can only generate a discrete and coarse measure of distance (i.e., they share the first k digits). Measures based on detailed firm-level information about patents, the type of labour employed, inputs, and other factors are appealing, but due to their data requirements, they can usually only be obtained at a relatively coarse level or only for a limited number of sectors.

Co-occurrence-based measures assume that if two products are produced or exported by the same unit (firm or country), there must be *something* needed to produce both products effectively and efficiently that is available within that unit. This feature could be specific factor endowments, knowledge, institutions, or technological capabilities. Most likely, these 'revealed similarity' measures represent a combination of these factors. These measures can be calculated for any classification system for which one can observe the occurrence of products across different units. The main problem with these measures is their interpretation: there is no certainty that what causes the co-occurrence is related to capabilities. Two products may co-occur because they have demand complementarities, like coffee pods and coffee machines.[6] In Section 4.4, alternative distance measures are used to precisely examine this possibility.

We build our analysis using a measure of distance that is heavily based on the measure of 'proximity' proposed by Hidalgo et al. (2007). With their proposed measure, the authors show that countries tend to evolve 'slowly' by developing goods that are 'close' to what they already are good at producing. Here, we will explore whether the same logic applies at the firm level.

3 Data and Definitions

3.1 Export data and definition of 'jumps'

The main data source for this study is a firm-product-level, six-digit Harmonised System database from Chilean customs for the years 1991-2006. The data contains information on over 5,000 firms and 2,000 types of export products (see Table 1). We complement this information with the

[6] However, even if demand complementarities ultimately drive co-occurrence, production capabilities are still required for these complementarities to co-occur in the same unit.

Table 1: Sample description, Chilean customs data 1991-2006.

Num. observations (jumps)	15179
Number of unique firms	5131
Number of unique products (HS6)	2208
Number of years	15

Notes. An observation is each case of a firm adding a previously unexported 6-digit HS code to its export basket.

Chilean manufacturing census to provide several robustness checks.[7]

With this database, we define a *firm-level new export product* (or 'jump') as an export code that has not previously been exported by a firm in our data.[8] In our database, this is indicated by a new HS6 code for a given firm. However, given that the HS classification is updated roughly every five years, we need a product classification that is consistent throughout the 1991-2006 period, seeking to eliminate the possibility of mistaking code changes for new exports. The methodology for concording the HS codes put forward by Wagner and Zahler (2015) is used.[9] This concording procedure is conservative in the sense that it eliminates the possibility of observing a 'new' product (and mistakenly observing failures) due to a change in classification, but at the same time, it may overstate the survival of a product if a substitution of one product for another is instead seen as a continuation of a broadly defined new product category. This would at the same time hide some cases of new exports.[10,11]

A limitation of customs data is that we cannot know with certainty whether the goods exported by a firm are produced by that firm or not. As we are interested in firms' underlying productive capabilities, we only want to include goods produced by the respective exporting firm. We address this and other issues by defining firm- and product-level filters similar to those used by Wagner and Zahler (2015):[12]

- Re-exports are dropped. Re-exports are defined as any product exported in year t that the same firm is importing in years t or $t-1$ for at least the same value at which the product is exported.

[7] We analyse and use mostly export data and not production data due to lack of availability of detailed micro-data for production for all sectors in the Chilean economy and because our motivation is to contribute to the understanding of the micro-level dynamics underlying the processes of export diversification. In this sense, we are assuming that a new export is also a new product for a firm, and this implies the availability of the capabilities and technologies required to produce and export the new product. However, since we do have access to production data for the manufacturing sector, we will use this data for some checks that support this interpretation.

[8] One obvious limitation is that for some firms, our data is left censored. As part of the robustness checks, we confirmed that the results hold when dropping the first few years of data (up to five years).

[9] The basic procedure of their methodology is to iteratively collapse codes into a 'minimum common code' that absorbs all codes that might be mistakenly classified as 'new' products. The code that is associated with a larger export value is left as the 'minimum common code'. For a more detailed explanation of the methodology, the reader is referred to Appendix A in Wagner and Zahler (2015).

[10] As part of robustness checks, we restrict the sample to periods during which there are no changes in classification

[11] From this point, any mention of HS refers to the concorded version of the HS codes.

[12] See also the discussion below about domestic sales data.

- Exports of less than USD 1,000 in their first year and USD 3,000 in total are dropped to avoid including export samples in our analysis.
- Firms that export (for the whole period) more than 25 different product categories are dropped as these are likely to be trade intermediaries. The cutoff was defined by manually examining the names of the firms with a high number of export products.
- There are unusual cases in which firms specialised in sectors different from machinery are found to export machinery, possibly for repairs. These incidents would appear in the data as relatively 'distant' products that are exported for only one year. To be on the safe side and avoid this source of bias, all 'jumps' toward products within the HS Section XVI (machinery) are dropped.[13]

In addition to trying to reduce the chance of considering firms that do not produce what they export, these filters also help reduce the risk of reverse causality due to short-lived exports that were predetermined as such for different reasons (samples, for instance) and might have systematically higher measured distances.[14]

We complement the use of the trade data as explained above with the use of the Chilean Manufacturing Census (ENIA) for 1995-2006. Doing so allows us to use data on actual firm production, albeit at the cost of analysing only the manufacturing sector. ENIA provides information on variables such as workers, total costs, and investment as well as a list of the products manufactured by each firm. We use this data for exercises that reinforce our interpretation of the distance measure, i.e., that the measure is capturing complementarities in production.

Next, we define survival (or duration, in survival analysis parlance) as the number of years between the first and last time a product is observed to be exported by a firm in our sample.[15]

Table 1 describes our sample of 'jumps' to new exports.

3.2 Defining *distance* between two exported products

We define the distance between two products using an approach that is heavily based on Hidalgo et al.'s (2007) definition of *proximity*.

These authors define the proximity between products i and j as:

$$\phi_{i,j} = min\left\{Pr\left[RCA_i > 1 | RCA_j > 1\right], Pr\left[RCA_j > 1 | RCA_i > 1\right]\right\} \quad (1)$$

where RCA is the revealed comparative advantage as defined by Balassa (1965):

$$RCA_{i,c} = \frac{x_{i,c} / \sum_i x_{i,c}}{\sum_c x_{i,c} / \sum_i \sum_c x_{i,c}}$$

[13] This is not a major issue as machinery represents a very small proportion of Chilean exports.

[14] The paper's findings are robust to broad changes in these filters.

[15] We tried the alternative definition of *continuous* duration: the number of years that a product is exported uninterruptedly by a firm after its first appearance. Both measures represent opposite and alternative definitions in terms of the time that we allow to pass in which there were no exports to deem that a spell had ended. The results (unreported) using *continuous* survival are very similar.

where $x_{i,c}$ are the exports of product i from country c.
The proximity between two products i and j is then the minimum of the probability of a country exporting i conditional on exporting j (both with RCA>1).[16]

With this definition, and using the BACI world trade flows database for the year 2002, we construct a 4614×4614 symmetric matrix $\{\phi_{i,j}\}$[17] *excluding* Chile to ensure that the proximity measures are exogenous to the decisions of Chilean exporters.

We define *distance* as $1-\phi_{i,j}$ in order to be closer to the idea of 'distance from the core' used in the literature. The expectation is that products that are more 'distant' from the current basket will be manufactured less efficiently and will therefore have a lower probability of 'surviving' a certain number of years. Examples of pairwise distances are shown in Appendix A.

Both the models of Eckel and Neary (2010) and Bernard et al. (2011) predict that firms will obtain lower revenues from products that are more 'distant' from the firm's core and that the profit margin will be smaller for more 'distant' products.[18] The latter implies that fringe products are more likely to be abandoned after a negative shock, as shown empirically by Iacovone and Javorcik (2010) for Mexican manufacturers. Following Eckel and Neary (2010), for each export 'jump' for a given firm, we define the *distance from the core* as the distance between the top-selling export product in $t-1$ and each new exported variety.[19] Thus, the distance between the firm's core and product p added by firm f in year t is:

$$dist\_core_{p,f,t} = 1 - \phi_{0,p}$$

where 0 indexes the firm's core product (firm f's product with the largest exports in $t-1$, the year before product p is introduced).[20] We assume that the core competence of a firm is best expressed in its top exported product at a given moment in time, as in Eckel and Neary (2010). As mentioned before, in the model in the work above there is one 'core competence', and all varieties are ordered along a single axis according to how distant they are from their core competence. However, as mentioned by Prahalad and Hamel (1990), firms can have multiple competences or capabilities on which to draw to produce and export different products. If, as posited by the latter authors, firms require different combinations of a number of competences for different goods, then the distance from the top-selling product will not necessarily capture the distance between a new product and the firm's existing capabilities. For instance, a product may be distant (in terms of competences) from the top product but very close to the second-highest

[16] Hidalgo et al. (2007) require revealed comparative advantage in a product to consider that there is co-occurrence instead of simply exporting it as a way to ensure that a country has the required endowments and capabilities and is an efficient producer of a good. Using the minimum of both conditional probabilities produces a symmetric measure and avoids the problems that would occur with measures such as the joint probability for products exported by a small number of countries. See Hausmann and Klinger (2007) for a discussion.

[17] 4614 is the number of HS classifications at the 6-digit level that are traded during 2002, which we chose arbitrarily among the years with the highest number of HS codes traded.

[18] Only Eckel and Neary (2010) explicitly refer to the 'core' and to a 'distance', although Bernard et al. (2011) can also be interpreted as firms having a core and different products being more or less distant from it.

[19] The theoretical predictions relating distance to revenues and profitability hold for domestic and foreign sales (see Eckel et al., 2016). Appendix B shows that the results that will be discussed in 11 this section hold for domestic production too, supporting the idea that our findings and our distance measure are related to the technological competences required to produce different goods.

[20] The core product depends on the firm and on time, but the product is simply indexed with a zero to make the notation cleaner.

selling product, for which the firm must also possess the required competences.

To empirically account for this possibility, we assume that firms are more specialised at producing the goods that they export at higher values, and we define a value-weighted average of the pairwise distances between each product in the basket in $t-1$. This process collapses the multiple possible underlying competences that are implicit when defining distance on a product-to-product basis into a unidimensional 'distance from the core'.[21]

The weighted distance between a new export and the firm's previous export basket is defined as:

$$weighted\ distance_{p,f,t} = \sum_{k \in \Theta_{f,t-1}} \omega_{f,k,t-1}(1 - \phi_{p,k}) \tag{2}$$

where p is the HS6 product code of the newly exported product, f the firm identifier, t is the year when the 'jump' occurs and k indexes the products exported by the firm in the previous period.

$\Theta_{f,t-1}$ is the set of HS6 codes exported by firm f in period $t-1$.

$\omega_{f,k,t-1}$ is the share of product k in the value of exports from firm f in period $t-1$ (sorted descending by sales).

$\phi_{p,k}$ is the proximity between HS6 codes p and k as defined in equation 1.

$weighted\ distance_{p,f,t}$ is then a value-weighted average of the pairwise distances between the new export and those products exported by the firm in the previous year. Both this value and that of $\phi_{p,k}$ lie between 0 and 1.

3.3 Descriptive statistics and preliminary evidence

Figure 1 plots the distribution of all the distances for every possible pair of products in the HS classification [22] and the distribution of the observed weighted distances of the 'jumps' in the data. The distribution of actual jumps has more variance, showing that firms 'jump' over a broad range of distances. Actual jumps tend to be much shorter than what a random jump would resemble. Table 2 shows some statistics on the jumps and unconditional survival. The average distance is 0.7, which is roughly equivalent to a firm to moving from *'tomatoes'* to *'tomato ketchup and other sauces'*, for example.[23] The average survival of a jump is 1.7 years, but the median jump does not survive beyond the first year of exports. Firms that jumped towards new products on average exported 2.9 (a median of 2) products before and had a median of \$370,000 of total exports per year. New products (in the year of introduction to export markets) represent a median of 6.7% of the value exported by the firm that year.

Next, we analyse our measure of distance. We want to test if *dist_core* predicts the revenues generated by and the survival of the products added by a firm to its basket–as predicted by

[21] An alternative measure could be the pairwise distance to the closest good in the firm's basket. However, using the shortest distance does not account for how the new good might require capabilities that are not needed for the closest good but that are needed for other goods that the firm exports. Moreover, firms develop certain capabilities more than others. If the good that is closest to the new export is a fringe good for the firm, using that distance would probably be misleading.

[22] These are the 10,642,191 unique off-diagonal terms from the matrix of distances.

[23] More examples are presented in Appendix A.

Table 2: Descriptive statistics for new firm-product export spells (selected variables)

	Mean	Median	Min	Max
Weighted distance	.707	.722	.161	1
Survival (years)	1.7	0	0	14
No. of products before jump	2.93	2	1	19
Volume initial year (USD)	141,287	15,360	1,000	3.76e+08
Total firm volume initial year (USD)	4,237,457	367,276	1,000	2.18e+09
Share of firm exports in first year	.215	.067	2.97e-06	1

Notes. Weighted distance is a value-weighted average of the pairwise distances between the new export product and each product that the firm was previously exporting. Survival is the number of years that a new firm-product is exported after the year in which it was introduced. No. of products before the jump is the number of products that the firm was exporting in the year before it added the new export.

theory–after controlling for the measures that have been used before as proxies for distance.[24]

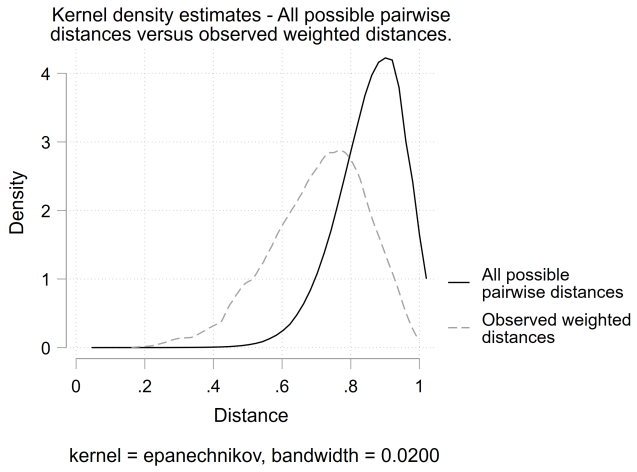

Figure 1: Distribution of all possible pairwise distances and observed weighted distances.

Table 3 presents the results of regressing the log exports of added products (in their first year) on the distance to the core measure defined above as well as the previously used proxies for distance. We also include six-digit HS code and year dummy variables. Table 4 analyses the number of years for which a new export survives, controlling for the same variables and additionally for the (log) initial value exported of the new product. The results for survival years also provide a preliminary look at the main question that we address in the paper.

The first column in each table includes only the distance from the core measure, and the following columns add the different proxies. *dist_core* is a significant determinant of sales and survival on top of each of these proxies for distance, suggesting that this variable carries

[24] Iacovone and Javorcik (2010) proxy for distance using a product's sales, the proportion of total sales that this value represents for the firm, and the share of the firm's sales over all domestic sales of that product. Eckel et al. (2015) use dummies ranking the products according to their sales.

Table 3: (log) Exports of new product (first year) and measures of distance.

Dep.var.: (log) exports	I	II	III	IV
Distance from the core	-0.864***	-0.646***	-0.716***	-0.548***
	(0.000)	(0.000)	(0.000)	(0.000)
Share of the firm's exports		3.347***		
		(0.000)		
Share of the country's exports			3.233***	
			(0.000)	
Ranking dummies				✓
N	12227	12227	12227	12227
R2	0.563	0.677	0.604	0.659

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is an addition of a new 6-digit HS code to a firm's export basket. The dependent variable is the log of exports of the new product during its first year. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Survival of new product (years) and measures of distance.

Dep.var.: years survived	I	II	III	IV	V
Distance from the core	-2.200***	-2.010***	-2.104***	-2.159***	-2.155***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(log) Initial value		0.220***			
		(0.000)			
Share of the firm's exports			1.492***		
			(0.000)		
Share of the country's exports				0.902***	
				(0.000)	
Ranking dummies					✓
N	12227	12227	12227	12227	12227
R2	0.656	0.663	0.663	0.657	0.657

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is an addition of a new 6-digit HS code to a firm's export basket. The dependent variable is the number of years that the new spell survives after its first appearance. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. (log) initial value is the (log) export value of the new product. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: (log) Exports of new product (first year) and measures of distance.

Dep.var.: (log) exports	I	II	III	IV
Distance from the core	-0.779***	-0.600***	-0.641***	-0.409***
	(0.000)	(0.000)	(0.000)	(0.004)
Weighted dist. from non-core	-1.181***	-0.662***	-1.059***	-1.806***
	(0.000)	(0.000)	(0.000)	(0.000)
Share of the firm's exports		3.319***		
		(0.000)		
Share of the country's exports			3.205***	
			(0.000)	
Ranking dummies				✓
N	12227	12227	12227	12227
R2	0.567	0.678	0.607	0.668

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is an addition of a new 6-digit HS code to a firm's export basket. The dependent variable is the log of exports of the new product during its first year. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. Weighted dist. from non-core is the weighted distance between the new product and all products exported in t-1 minus the distance from the core. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level. p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

information that is not captured by the proxies.[25]

Next, we show that the *weighted* measure of distance seems to include additional information beyond that captured by *dist_core*, supporting the argument that firms have a number of different 'core competences'. Tables 5 and 6 replicate the results for export value and survival from Tables 3 and 4, adding a measure of distance to all products *except* the top-selling one. This measure is simply *weighted distance* as defined above but subtracting the contribution of the top selling product:

$$weighted\ distance\ noncore_{p,f,t} = weighted\ distance_{p,f,t} - \omega_{f,0,t-1}(1 - \phi_{0,p})$$

As seen in the tables, the 'noncore' distance measure is significant on top of the 'distance from the core' used before (and on top of the previously used proxies for distance), indicating that the rest of the export basket, beyond the top selling product, also help explain the sales and survival of new exports.[26]

[25] To compare the explanatory power of the different measures, similar regressions were conducted but including each proxy (and our measure) separately (these results are not reported). The R^2*s* indicate that our distance measure has a lower explanatory power for sales–which is not surprising given that the other proxies carry information about sales. However, when performing the same analysis for survival, the explanatory power of our distance measure is roughly the same. More precisely, the R^2 is slightly smaller for two of the other proxies and is slightly higher for the other two (one of which is the set of ranking dummies, for which both the adjusted and non-adjusted R^2s are higher for our measure).

[26] The coefficients between the two measures should not be compared as the non-core variable is mechanically smaller because the shares do not add up to one.

Table 6: Survival of new product (years) and measures of distance.

Dep.var.: years survived	I	II	III	IV	V
Distance from the core	-2.039***	-1.880***	-1.964***	-2.003***	-1.971***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Weighted dist. from non-core	-2.237***	-1.994***	-2.017***	-2.205***	-2.382***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(log) Initial value		0.205***			
		(0.000)			
Share of the firm's exports			1.407***		
			(0.000)		
Share of the country's exports				0.843***	
				(0.000)	
Ranking dummies					✓
N	12227	12227	12227	12227	12227
R2	0.660	0.666	0.667	0.661	0.662

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is an addition of a new 6-digit HS code to a firm's export basket. The dependent variable is the number of years that the new spell survives after its first appearance. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. Weighted dist. from non-core is the weighted distance between the new product and all products exported in t-1 minus the distance from the core. (log) initial value is the (log) export value of the new product. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4 Results

To study whether new export products that are 'closer' to the firm's previous export basket are more likely to succeed, the dependent variable could be the number of years for which the new export 'survives'. Since export survival is likely to be right censored, we will complement the linear models with survival analysis. In addition to correctly modelling the nature of the variable of interest, these models deal explicitly with right censoring (the products that survive until the end of the sample). The empirical literature that looks at the survival of trade flows has evolved into increasingly flexible and appropriate models, which are discussed in Appendix C.

A first simple view of using these methodologies is shown in Figure 2, which plots the Kaplan-Meier survival function for 'jumps' below and above the lower and upper quartiles. An unreported log-rank test shows that the functions are significantly different; however; most of the difference is in different survival rates in the first year, i.e., whether the product survives at all or is dropped after its first appearance. This is consistent with the idea of experimenting to learn about profitability, as in Albornoz et al. (2012).

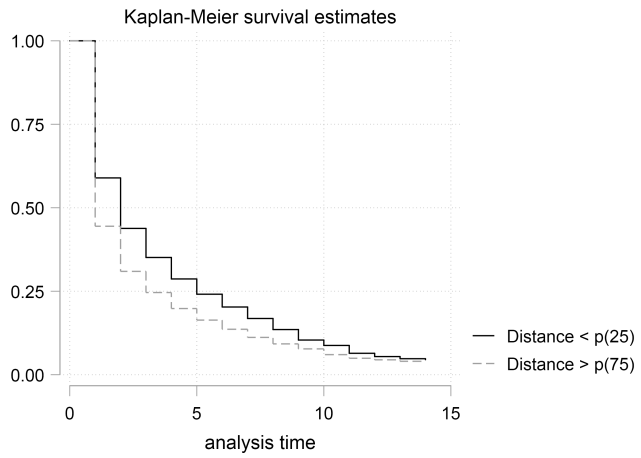

Figure 2: Kaplan-Meier estimator of unconditional survival (probability of surviving up to year t) for jumps above and below the upper and lower quartiles of distance. See Appendix C for more details.

4.1 Fixed effects

We begin our main econometric analysis using a linear probability model with fixed effects, with the goal of analysing different and increasingly demanding sources of unobserved heterogeneity,

which could be correlated with the impact of distance on firm-product survival[27]. From now on, we standardise weighted distance (mean zero, unitary standard deviation) to simplify the interpretation of our results.

As the main regressor of interest is time-invariant,[28] and to facilitate the comparison of partial effects with the nonlinear survival models presented later, we set up the data as a cross-section in which each record represents a 'jump'. Equation (3) shows the basic regression model.

$$survives_{p,f,t} = survives_i = \beta distance_i + \delta' \mathbf{x}_i + \lambda_p + \lambda_f + \lambda_t + u_i \tag{3}$$

The dependent variable $survives_i$ is a binary variable indicating whether jump i (given by firm f to product p in period t) survived for at least one year after its initial appearance (in other words, if the firm exported the product during two or more years). The previous result in figure 2 justifies this definition.

The controls in $\mathbf{x}_i$ are measured at the moment of the jump or before, and λ_p, λ_f and λ_t represent the product, firm and year-of-jump fixed effects, respectively.

Table 7 presents increasingly demanding specifications, showing that the relationship between distance and survival remains significantly negative even when controlling for firm-year and product-year fixed effects simultaneously in the last column. Standard errors are clustered at the level of four-digit HS sectors to account for possible correlation across the residuals that might remain after controlling for firm and six-digit product fixed effects.

Model I includes controls for the number of products exported the year before the jump, the initial share of the new export of the firm's total exports (in the year that the product was introduced), firm age, the (log) initial exported value of the new product, (log) RCA, growth of total firm exports, and price premium.[29] Models II to VI drop some of these controls because they become collinear with some of the fixed effects. Model I includes dummy variables only for the year of the jump. Column II adds firm dummies that control for time-invariant firm-level unobserved heterogeneity (for example, managerial capability and the propensity of a firm to take risks, if these are assumed to remain constant throughout the period). Column III adds dummies for six-digit HS codes. Each product may have an intrinsic degree of complexity that is beyond what is captured by the distance measure and that makes it more or less difficult to export successfully. If this intrinsic complexity is also correlated to distance, such a relationship would bias our estimates. Columns IV and V go one step further by controlling for time-varying firm- or product-level unobserved heterogeneity. Firm-year dummies control for idiosyncratic, firm-level shocks, and product-year dummies control for issues like shocks to world prices, shocks

[27] We will compare the results with survival models. If there is dependence between the spells, for instance between jumps by the same firms or between jumps to the same product by different firms, the survival models would be misspecified. In a linear setting, fixed effects can be used to solve this problem, but in nonlinear models, fixed effects can only be included for groups with a large number of observations; if fixed effects for groups with a small number of observations were included (such as firms and products), all the estimated coefficients would be biased Wooldridge (2010).

[28] We only define distance relative to the basket for the year right before the jump, and we do so for two reasons: first, as distance depends on the shares of different exports, it would be highly endogenous to survival if it changed with time, and second, we are interested in the firm's competences at the moment at which the product was added to its export basket.

[29] Defined as the log of the firm's unit value for the product over the yearly value-weighted average unit value of all domestic exporters of the product.

Table 7: Determinants of survival of new firm-level product exports.

Dep.var.: survival>0 years	I	II	III	IV	V	VI
Weighted distance	-0.0484***	-0.0516***	-0.0581***	-0.0634***	-0.0466***	-0.0497*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.060)
No. prod. before jump	0.000199	-0.0340***	-0.0305***		-0.0255***	
	(0.935)	(0.000)	(0.000)		(0.000)	
Initial share new prod	-0.0185	0.157***	0.166***	0.0239	0.150***	0.0503
	(0.475)	(0.000)	(0.000)	(0.637)	(0.000)	(0.475)
Firm age	0.0000756	-0.0774***	-2.022		-1.565	
	(0.968)	(0.000)	(1.000)		(1.000)	
Ln(Initial value)	0.0232***	0.0301***	0.0375***	0.0567***	0.0412***	0.0463***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln(Total firm exports)	0.0137***					
	(0.003)					
ln(RCA)	-0.00121	-0.00517	-0.0398***	-0.0259*		
	(0.755)	(0.235)	(0.000)	(0.072)		
Log-diff of firm exports	0.0274***	0.00334	0.00474		-0.00151	
	(0.000)	(0.618)	(0.466)		(0.897)	
Price premium	-0.00954**	0.00393	-0.00111	0.00208	0.00267	0.0100
	(0.020)	(0.485)	(0.859)	(0.806)	(0.798)	(0.682)
No. others exporting	0.0169***	0.0443***	-0.0113	-0.0113		
	(0.006)	(0.000)	(0.579)	(0.703)		
Firm dummies		✓	✓		✓	
HS dummies			✓	✓		
Year dummies	✓	✓	✓			
Firm-year dummies				✓		✓
HS-year dummies					✓	✓
N	10759	10759	8297	5291	5204	2988
R2	0.132	0.557	0.581	0.711	0.687	0.793

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is a case of a firm adding a previously unexported 6-digit HS code to its export basket. The dependent variable is a dummy equal to one if the new export spell survived for at least one year after its initial appearance. Weighted distance is the (value) weighted distance between the new product and all products exported in t-1. No. prod. before jump is the number of products the firm was exporting in t-1; Initial share new prod is the (value) weight of the new product in the total value exported by the firm in t; firm age is the age of the firm in the database; (log) initial value is the (log) value of the new exported product in t; (log) Total firm exports is the (log) value of all products exported by the firm in t; (log) RCA is the (log) RCA of the new product at the country level, in t; Log-diff of firm exports, is the log difference of total exports of the firm from t-1 to t; Price premium is the ratio of the unit value exported of the new product to the average unit value of the same product at the country level; No. others exporting is the number of other firms exporting the new product in t. Standard errors clustered at the 4-digit HS level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

to demand for a product, and changes in quality or other characteristics within the product category that could be related to distance and survival. Finally, column VI includes both the firm-year and HS code-year sets of dummies. In this case, the coefficient is only weakly significant, but this result is found under a very demanding specification and a smaller sample size.

It is difficult to think of omitted factors that could be correlated to distance and survival and that are not accounted for in these specifications.

The weighted distance to the firm's previous export basket has a significantly negative coefficient under all specifications. This is the main result from the paper: *new exports that are relatively more distant to the firm's previous exports face a higher risk of 'dying'.*

The coefficients are relatively stable, even under very demanding specifications. Model III is our preferred specification, given that the estimate for weighted distance lies between those from the more demanding specifications of models IV through VI. The results indicate that a one standard deviation longer distance of a jump decreases the probability of survival of the new export beyond the first year by approximately six percentage points, which is a 12% drop relative to the unconditional probability of surviving for at least one year (around 50%). A difference of one standard deviation between two jumps (approximately 0.2 in our distance measure) is a common occurrence. For example, the distance between *'wine (not sparkling)'* and *'wine (sparkling)'* is 0.56, and the distance between *'wine (not sparkling)'* and *'other fermented beverages'* is 0.75. Thus, if a firm is exporting wine (not sparkling) and jumps to a new export, the probability that sparkling wine survives past the first year is 12% higher than for other fermented beverages.

The relationship between survival and the included covariates is as expected. The number of products that the firm was exporting before is negative (consistent with Eckel and Neary, 2010), and the initial value and share for the firm of the new export is positive (consistent with Rauch and Watson, 2003). These results are also consistent with the empirical literature reviewed above except for the insignificance of some controls that had been found to be significant, like the price premium[30]. The fact that RCA is not a strong determinant of survival gives us a hint that the issue with product survival for a firm does not appear to be associated with country level or sector-level development, as compared to firm-level characteristics.

4.2 Survival models

The theory behind survival analysis is discussed in Appendix C. The hazard function defines the probability that a new export 'dies' during a certain period k, conditional on having survived up to that period. For a 'jump' towards product p given by firm f in year t, the hazard takes the following form:

$$h_{p,f,t,k} = h_{i,k} = F\left(\beta distance_i + \gamma_k + \lambda_t + \delta' \mathbf{x}_{i,k}\right) \tag{4}$$

[30] This is defined as the log of the firm's unit value for the product over the yearly value-weighted average unit value of all domestic exporters of the product.

where p is the HS6 product code of the newly exported product and k indexes the number of periods for which a jump has survived.

To simplify the notation, the 'jump' to product p performed by firm f in period t is simply indexed as jump i; $h_{i,k}$ is the hazard rate of jump i in its kth year; $F(.)$ is a cumulative distribution function that defines the hazard; $distance_i$ is the the weighted distance between the new product and the products exported by the firm in the previous year, as defined in equation 2 in Section 3.2; γ_k is the nonparametric baseline hazard for each period; and λ_t represents the dummies for the starting year of each spell[31]. The set of controls in $\mathbf{x}_{i,k}$ includes variables such as the initial value of the trade flow and Chile's revealed comparative advantage in a product each year. The linear index that defines the argument of $F(.)$ is very similar to the OLS specifications with the difference that firm-level fixed-effects should not be included.

Table 8 reports the results for the cloglog, logit and probit models for the hazard function both with and without random effects. As is usual with nonlinear models, the coefficients are not directly comparable across models. The table reports the coefficients only so that the signs and significance levels can be examined.

[31] These variables are included to ensure exogeneity of the right censoring.

Table 8: Survival models —Hazard function parameter estimates, for different functional forms for the hazard.

Dep.var.: Pr(die|survived up to k)	Clog-log	Logit	Probit	RE Clog-log	RE Logit	RE Probit
Weighted distance	0.113***	0.147***	0.0881***	0.296***	0.360***	0.203***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(log) Initial value	-0.00995	-0.0140	-0.00815	-0.0778***	-0.0947***	-0.0534***
	(0.215)	(0.173)	(0.182)	(0.000)	(0.000)	(0.000)
(log) RCA	-0.00840	-0.0188**	-0.0135**	0.0104	0.0118	0.00649
	(0.229)	(0.038)	(0.013)	(0.484)	(0.527)	(0.537)
No. prods prejump	0.0587***	0.0760***	0.0467***	0.138***	0.167***	0.0944***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
No. prods exported	-0.0305***	-0.0369***	-0.0225***	-0.0756***	-0.0906***	-0.0512***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(log) Total firm exports	-0.0897***	-0.0999***	-0.0575***	-0.256***	-0.315***	-0.178***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log-diff of firm growth	-0.127***	-0.181***	-0.108***	-0.125***	-0.181***	-0.104***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Price premium	0.0189*	0.0253**	0.0150*	0.0328*	0.0421*	0.0241*
	(0.054)	(0.049)	(0.051)	(0.067)	(0.063)	(0.061)
Initial share new prod	-0.367***	-0.398***	-0.231***	-0.830***	-1.023***	-0.579***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
New 4-digit sector	0.220***	0.294***	0.178***	0.467***	0.567***	0.320***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
No. others exporting	0.0172*	0.0393***	0.0266***	-0.0399*	-0.0436	-0.0241
	(0.088)	(0.002)	(0.001)	(0.085)	(0.128)	(0.137)
Jump year dummies	Yes	Yes	Yes	Yes	Yes	Yes
HS2 dummies	Yes	Yes	Yes	Yes	Yes	Yes
Sample size	28572	28572	28572	28572	28572	28572
Log-pseudolikelihood	-17025.8	-17025.2	-17025.7	-16798.7	-16776.4	-16774.9
Rho=0 (p-value)				0.000	0.000	0.000

Notes. Nonlinear survival models, table reports the estimated coefficients for the hazard functions (more details in Appendix C). Chilean customs data (1991-2006). Weighted distance is the (value) weighted distance between the new product and all products exported in t-1. (log) initial value is the (log) value of the new exported product in t; (log) RCA is the (log) RCA of the new product at the country level, in t; No. prod. prejump is the number of products the firm was exporting in t-1; No. prods exported is the number of products a firm exports each year; (log) Total firm exports is the (log) value of all products exported by the firm in t; Log-diff of firm exports, is the log difference of total exports of the firm from t-1 to t; Price premium is the ratio of the unit value exported of the new product to the average unit value of the same product at the country level; Initial share new prod is the (value) weight of the new product in the total value exported by the firm in t; New 4-digit sector is a dummy indicating that a product is new to the firm in its fourth digit; No. others exporting is the number of other firms exporting the new product in t. All models include dummies for a non-parametric baseline hazard function. Robust standard errors in case of no dynamic completeness (see Wooldridge, 2010). The Rho=0 p is the p-value for a LR test of whether there are no random effects. It is based on the same regression but without robust standard errors, to make the LR test meaningful.

p-values in parentheses. * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01.

The first thing to note is that for the three models, the hypothesis that the proportion of the variance contributed by the random effects is zero is strongly rejected (see the $\rho = 0$ test). Hence, the models with unobserved heterogeneity should be preferred. Appendix D reports the semi-elasticities $\partial ln(h_{it})/\partial distance$ for the random effects models evaluated at different durations.

As other empirical studies have found, there are no major differences across models in the quality of the fit, as seen by the similar log-likelihoods. The probit model may be preferable because of the substantial degree of non-proportionality in its hazards (see Appendix C for a discussion).

The main result from the previous section–that *distance increases the risk* of a new export–holds strongly under the survival models, as shown by the significance and the sign of the coefficients on distance (positive indicates a higher hazard). As firm and HS code fixed effects cannot be included in these models, there are more covariates across all specifications. The results for the controls are consistent with the linear models, and some coefficients that were not significant or that were collinear with the fixed effects are now significant.

Figure 3 presents the estimated survival function (the unconditional probability of surviving up to t years) for the typical 'long' and 'short' jumps[32], for both the cloglog and the probit specifications for the hazard. Just like the Kaplan-Meier estimators in Figure 2 suggested, distance is much more relevant in determining whether the new firm-product pair survives its first and second years, and the importance of distance fades as a product survives for a longer period of time.

In Table 9, we allow the coefficient for distance for the first year that the new product is exported to differ from the rest of the periods. Both the coefficients for the interaction term and for the original distance variable are positive and strongly significant. This result confirms that the relationship is different for the first period—beyond what the nonlinearity in the model allows for—but also that distance is still associated with an increased hazard beyond the first period.

Are these effects economically meaningful? It is possible to examine the semi-elasticities ($\partial ln(h_{it})/\partial distance$) evaluated at the modes and means of the covariates (see footnote 32). The approximate proportional changes in the probabilities of survival in the first period, when the weighted distance increases by one percentage point (in the models with an interaction term for the first period), are 0.294, 0.317 and 0.296 for the cloglog, logit and probit models with random effects, and for the second period, these values drop to 0.196, 0.177 and 0.158, respectively (see Table D1). This result means that for a 'typical' jump, a difference in distance of one standard deviation translates into roughly a 30% change in the probability that the new spell stops after its first year and a 16-20% change in the probability that the new spell stops after its second year. The equivalent marginal effects in the model without the interaction–where the differences in the effects across periods are exclusively the result of the nonlinearity–are approximately 26% and 23%. The model manages to capture some of the difference in the hazards in the first

[32]The distance for the 'short' jump is evaluated at the first quartile of its distribution, while for the 'long' jump, distance is evaluated at the third quartile. All covariates are evaluated at their means, except for the new 4-digit sector, jump year and HS2 dummies, evaluated at their modes.

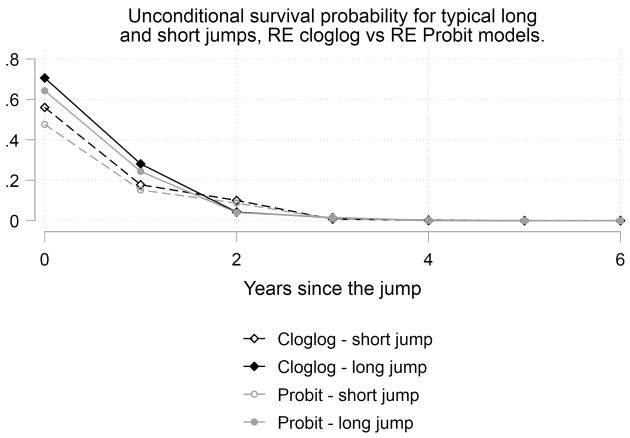

Figure 3: Estimated survival functions for the typical 'short' and long 'jumps', cloglog and probit models with random effects. The distance for the 'short' jump is evaluated at the first quartile of its distribution, while for the 'long' jump, distance is evaluated at the third quartile. All covariates are evaluated at their means, except for the new 4-digit sector, jump year and HS2 dummies, evaluated at their modes.

periods, but does not capture these differences as well as when the interaction term is included (compare Tables D1 and D2).

The results presented above are evaluated at a typical' jump. To be able to directly compare these estimates with those from the regressions with fixed effects, Table 10 reports the marginal effects of weighted distance on the probability of surviving at least one year for each random-effects model in Table 8 (for a typical jump). The marginal effects from the RE cloglog, logit and probit models are 0.0523, 0.0639 and 0.0641, respectively. These are very close to the coefficients for weighted distance in Table 7: -0.0634 and -0.0466 in the most demanding specifications and -0.0581 in our preferred specification (Model III). Considering that the linear models allow us to control for the firm and six-digit HS code fixed effects, our preferred specification for the

Table 9: Survival models with an interaction between distance and first period at risk.

Dep.var.: Pr(die|survived up to k)	RE Clog-log	RE Logit	RE Probit
Weighted distance	0.246***	0.275***	0.153***
	(0.000)	(0.000)	(0.000)
Weighted distance in first period at risk	0.0831**	0.141***	0.0828***
	(0.010)	(0.001)	(0.001)
Sample size	28572	28572	28572

Notes. Nonlinear survival models, table reports the estimated coefficients for the hazard functions (more details in Appendix C). Chilean customs data (1991-2006). The estimations include the same controls as in Table 8, although only the coefficients for weighted distance are reported. Weighted distance is the (value) weighted distance between the new product and all products exported in t-1. All models include dummies for a non-parametric baseline hazard function. Robust standard errors in case of no dynamic completeness (see Wooldridge, 2010). p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: d(h)/d(Weighted distance) at survival=0 for random effects models

	RE Clog-log	RE Logit	RE Probit
Weighted distance	0.0523***	0.0639***	0.0641***
	(0.000)	(0.000)	(0.000)
N	28572	28572	28572

Notes. Marginal effects of weighted distance on the probability of surviving at least one year at different survival times for each random-effects model in Table 8, for a typical jump (all covariates are evaluated at their means, except for the new 4-digit sector, jump year and HS2 dummies, evaluated at their modes.).
p-values in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

rest of the paper is still the linear model III from Table 7.

4.3 Robustness checks

Table 11 presents a battery of robustness checks for the main result. Given the similarity in the marginal effects between the linear and nonlinear models, the results presented here are for a linear model with a survival dummy as the dependent variable, as in Section 4.1.[33] All but the last column replicate our preferred specification (column III in Table 7) with different restrictions on the sample. Only the coefficients for the distance measures are displayed.

Models (1) and (2) drop the upper and lower tails of the distance measure to ensure that the results are not caused by a small number of extreme observations. Compared to the -0.058 in the original regression, the estimates barely change and only change in model (2).

Survival models naturally deal with censored observations. However, as an additional check in a linear setting, specification (3) restricts the sample to spells that are not right-censored. Distance is still strongly significant.

Model (4) considers the possibility that a 'new' product is not really new but rather the result of a mistake in the reported classification. In those cases, the firm would drop and add a product at the same time, and the count of the exported varieties would remain constant. To analyse this, we restricted the sample to include those firm-years where the increase in the number of distinct products exported is equal to or larger than the number of 'jumps' observed to filter out those false new products. Again, the results hold, albeit with a smaller coefficient.

To discard problems with the concordance procedure, model (5) restricts the sample to the 2002-2006 period, where the original data using a single HS classification are used. Again, the results hold.

Model (6) attempts to deal with the risk of misreporting as well as as that of problems with the concordance across HS versions. Mistakes in reporting are more likely in the last digit, as category labels sometimes differ only in a number or a single word. The sample is restricted to jumps that were new to the firm in their fourth HS digit (instead of only the sixth digit, as before), and the results hold.

Model (7) is perhaps the most demanding test. As shown in Figures 2 and 3, the hazard rates of short and long jumps primarily differ in the first year. One possible interpretation of this

[33] This variable takes a value of one if the new export survived for at least one year after its initial appearance.

Table 11: Robustness estimations for export survival and distance.

Dep.var.: Survival dummies	(1) distance<1.5	(2) distance>-2	(3) Drop censored	(4) Strictly adds
Weighted distance	-0.0580***	-0.0611***	-0.0589***	-0.0665***
	(0.000)	(0.000)	(0.000)	(0.000)
N	7925	7973	5679	3658
R2	0.584	0.582	0.584	0.647

Dep.var.: Survival dummies	(5) New 4-digit	(6) 2002-2006	(7) Survival>1	(8) Alternative distance
Weighted distance	-0.0451***	-0.0666***	-0.0188	
	(0.000)	(0.000)	(0.185)	
Alternative distance				-0.0207***
				(0.003)
N	5371	2918	3245	10759
R2	0.637	0.663	0.688	0.673

Notes. OLS regressions using Chilean customs data (1991-2006). Each observation is a case of a firm adding a previously unexported 6-digit HS code to its export basket. The dependent variable is a dummy that is equal to one if the new export spell survived for at least one year after its initial appearance. The only exception is model (7), where the dependent variable is a dummy that is equal to one if the new export spell survived for at least two years after its first appearance, conditional on having survived at least one.

Controls as in model III in Table 7 (year, firm and product dummies, as well as no. prods before jump, initial share new prod, firm age,(log) Initial value, (log) Total firm exports, (log) RCA, Log-diff of firm exports, price premium and no. others exporting). Standard errors clustered at the 4-digit HS level.

Models (1) and (2) restrict the sample by dropping extreme values of weighted distance. Model (3) drops right-censored observations. Model (4) includes only 'jumps' during firm-years where the change in the number of products exported by a firm is equal or larger than the number of jumps to new products. Model (5) includes only 'jumps' where the new product is new to the firm in its fourth digit (instead of only new in the fifth and sixth digits). Model (6) restricts the sample to the 2002-2006 period, where the same version of the HS classification is used. In model (7) the dependent variable is a dummy indicating whether the product survived for at least two years after its first appearance, conditional on having survived for at least one. Model (8) uses an alternative distance measure based on the digit of the new export that is new for the firm.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

is that firms are quick to learn whether new exports are profitable; thus, the effect of distance on survival is most pronounced in the first year of exporting. Alternatively, if the filters used for defining new products are insufficient, and some 'distant' products are a priori determined to be exported only once, the results in Tables 7 and 8 could suffer from reverse causality. For this test, the sample is restricted to products that survived for at least one year after their first appearances (they were not one-off exports), and the dependent variable is redefined as a dummy variable that equals one if the product survived for at least two years (in other words, exports stopped in year 3 or beyond). The coefficient drops to approximately a third of its original value and is no longer significant.

However, this happens estimating a model that absorbs more than 2,000 firm and 1,000 HS-code fixed effects with only 5,000 observations. Controlling for only firm-level *or* product-level unobserved heterogeneity (still clustering errors at the 4-digit HS code level), the estimated coefficients are closer to those seen in Table 7 and are significant at the 1% level. The changes in the coefficients across models I-III in Table 7 suggest that omitting the firm or the sector

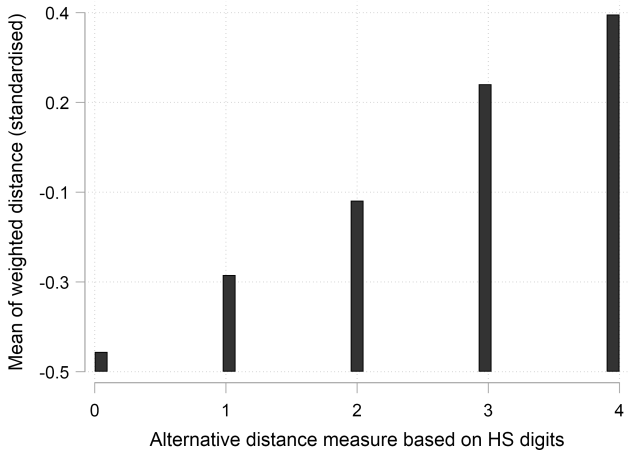

Figure 4: Average standardised weighted distance for each value of the alternative distance measure (discrete variable, value depends on the digit in which the product is new to the firm).

fixed effects does not cause much bias, and in this case, with a smaller sample, when those fixed effects are not included simultaneously, enough variability remains in the covariates to precisely estimate the parameters. Moreover, the smaller magnitude of the coefficient for survival after the first period at risk is consistent with the results for survival models with an interaction between distance and the first period (see Table 9). In other words, most of the effect of distance on survival takes place during the first two years of exports. This finding makes sense as a firm should realize whether its exports are competitive in a short time.

Finally, model (8) presents the results for an alternative measure of distance based exclusively on the HS codes. Assuming that products within a five-digit HS category (only differing in their sixth digit) are more closely related than products within a four-digit category (differing in the last two digits), it is possible to define a categorical variable that is increasing in distance. The maximum value of this variable is 4, representing when the new product's first HS digit is new to the firm; the variable takes a value of 3 if the firm previously exported something with the same first digit, but the second digit is new; the variable takes a value of 2 if only the third digit is new; the variable is 1 if only the fourth digit is new; and the variable is 0 otherwise. Figure 4 shows that this alternative distance measure is positively correlated to the original weighted distance measure. Both measures seem to be measuring something similar—an underlying 'similarity' between products. The alternative distance measure is strongly significant.[34]

4.4 Supply or demand complementarities?

One possible critique of the measure of similarity based on co-occurrence is that this approach does not say anything about *what* makes products *'similar'*. Two products might be co-exported

[34] A number of additional robustness tests were conducted but are omitted here for succinctness. These tests include using a non-weighted average distance to the basket or the distance to the closest good and dropping the first few years of 'jumps' (which might be re-entries), among others.

because they are complementary in terms of the capabilities required to produce them—as suggested by Hidalgo et al. (2007)—or because there are demand complementarities associated with them. To distinguish between these two scenarios, we use the measures of input and output similarity defined by Boehm et al. (2016). These measures define sectors (or products) to be similar in their inputs if the expenditure shares on inputs from other sectors are highly correlated. To consider the expenditure shares in all inputs simultaneously, the authors build a measure based on the inner product of the vectors with the shares of inputs sourced from each sector. For output similarity, the measure is based on the shares of the output of the sector sold to every other sector. The discussion below is about 'sectors', referring to the group of all firms that are producing within a certain category.

Defining $\omega_{i,j}$ as the share of sector i's intermediate inputs coming from sector j, input similarity is defined as the normalised inner product of the vectors of shares $\omega_i = (\omega_{i,1}, \omega_{i,2}, ...)'$ for sectors i and j:

$$input\ Similarity_{i,j} = \frac{\sum_{k=1}^{N} \omega_{i,k}\omega_{j,k}}{\sqrt{\sum_{k=1}^{N} \omega_{i,k}^2 \times \sum_{k=1}^{N} \omega_{j,k}^2}}$$

The numerator is the inner product of the vectors of shares. The denominator simply normalises the measure to be between zero and one. This expression is maximised when $\omega_{i,k} = \omega_{j,k}\ \forall\ k$, that is, when the structure of intermediate input usage by two sectors is identical, and the expression equals zero when the two sectors do not have any inputs in common.

The measure of output similarity is analogous but is based on the share of sector i's output bought by sector j, $\mu_{i,j}$:

$$output\ Similarity_{i,j} = \frac{\sum_{k=1}^{N} \mu_{i,k}\mu_{j,k}}{\sqrt{\sum_{k=1}^{N} \mu_{i,k}^2 \times \sum_{k=1}^{N} \mu_{j,k}^2}}$$

So, the above is a measure of how similar the sets of buyers from two sectors are (considering only sales for use as intermediate inputs). This measure takes a value of zero when the two sectors do not have any buyers in common and is one when the way in which their sales are distributed across other sectors is identical.

We interpret input similarity as a measure of technological complementarities between products and output similarity as a measure of market complementarities between the products.

It is possible to obtain the information required to build these measures from the ENIA manufacturing census, or more precisely, from two appendices to the survey that contain detailed information about the products manufactured by a plant and the intermediate inputs consumed (classified using the Central Product Classification (CPC) with five digits). The main shortcoming of this dataset is that when a plant manufactures more than one product, only total inputs are reported, and it is not possible to know how much of an input was allocated to the manufacture of each product. To avoid any imputations, only the data from single-product firms is used to obtain the shares of inputs used.[35]

[35] With single-product defined as plants that are active at only one five-digit CPC sector. These products represent 48% of the sales in the sample.

Table 12: Input- and output-based distance measures and survival

Dep.var.: survival>0 years	Survival>0	Survival>0
Weighted output distance	0.104	
	(0.558)	
Weighted input distance		-0.253***
		(0.000)
Sample size	3748	9154

Notes. OLS regressions using Chilean manufacturing data (1995-2006). Each observation is a case of a firm adding a new 5-digit CPC code to its production basket.

The dependent variable is a dummy equal to one if the new product survived for at least one year after the first year it was manufactured. Controls include year, firm and product dummies, as well as number of products being manufactured before the , the initial share of the new product in the firm's basket, firm age in the database, (log) initial sales of the new product, (log) plant employment, (log) TFP (estimated using the methodology by Levinsohn and Petrin, 2003), (log) capital of the firm and (log) investment, all in the year of the jump.

Standard errors clustered at the 5-digit CPC code level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

To define the share measures at the sector level, sales are aggregated over the 2001-2003 period[36] for every supplier-to-buyer pair of sectors, for instance, the total expenditure on bottles and corks by all wine makers and the total expenditure on fabrics and dyes by manufacturers of shirts. From this data, which includes all the sales of each product as an intermediate input and the total value of each input used to manufacture them, it is possible to obtain the required shares. Then, the vectors of shares supplied ω_i and shares bought μ_i are used to construct the measures of input and output similarity between each pair of sectors. Finally, just like before, a value-weighted measure of *distance from the basket* is obtained, as defined by equation 2 but replacing *proximity* ($\phi_{i,j}$) with the two similarity measures defined here. Here 'jumps' are defined as five-digit CPC categories added to the manufacturing plant's product mix, so we are looking at manufacturing, not exports.[37]

For succinctness, Table 12 presents only the coefficients for the distance measures from the estimation of the preferred fixed-effects specification, Model III from Table 7. Only the measure of distance in inputs is a significant determinant of the survival of new exports.[38]

This result suggests that it is not complementarities in demand that underlie the results from the previous sections but rather some form of economies of scope associated with similarity in the technological capabilities required to process the inputs and manufacture the products.

Moreover, this result reinforces the idea that the relationship between distance and survival is about *productive* capabilities and not about the capabilities required to successfully sell products manufactured by someone else to foreign markets.[39] This idea is strengthened by the results in Appendix B for domestic manufacturing sales.

[36] Data on input usage was only available for this period.

[37] For this reason, we include all 'jumps' to new products without applying the filters that we used for customs data.

[38] The difference in sample sizes is due to the number of sectors for which it was possible to define each similarity measure. When restricting the sample randomly to be of the same magnitude for both regressions, the coefficient for similarity in inputs is still strongly significant and larger in magnitude.

[39] The customs data has no information about who produced a good. However, the exercise presented in the section is based exclusively on information about manufacturing.

Overall, these results support the view that the baseline measure of distance used in this paper can be interpreted as capturing the distance between the competences or capabilities required to produce a certain new good, and the firm's 'core' competences. The advantage of that baseline measure vis-à-vis those defined in this section is that it can be defined for all goods because it does not require the detailed data on input usage that are only available for the manufacturing sector.

5 Summary

An important part of the growth in country-level exports is explained by new export varieties added by existing firms. Previous studies have shown that the survival of new export products, at the level of the firm, depends on factors such as the initial value of the trade flow and the firm's experience.

This paper presents evidence of a robust positive relationship between how 'distant' a new export is from the firm's current export basket and its hazard rate. This finding is consistent with both Eckel and Neary's (2010) and Bernard et al.'s (2011) models of multi-product firms, which suggest that the first products that a firm should stop exporting after suffering a negative shock are those that are more 'distant' from the firm's core activities. The measure proposed here can be interpreted as a proxy for the theoretical 'distance from the core', which additionally takes into account the fact that different underlying 'competences' are required for manufacturing different goods.

We first show that the distance between a new product and the firm's 'core' (represented by its top selling export) is a significant determinant of export value and survival, as predicted by theory. We also provide evidence suggesting that not only the top-selling product matters but that firms have multiple competences that are combined to produce different goods, as proposed originally by Prahalad and Hamel (1990). Based on this, we define a distance measure that considers the whole export basket and incorporate the measure into an empirical survival analysis exercise.

Using data from Chilean exporters and different econometric techniques (linear models and survival analysis), the paper shows that when a firm adds to its basket an export that is further away from its current capabilities, the new export product has a higher risk of failure, especially during its first year. More specifically, a one standard deviation larger distance is associated with a 12% change in the probability of surviving beyond the first year.

To understand *what* is being captured by the distance variable, alternative distance measures that attempt to capture complementarities in technology and in market demand were used. Only the measure of technological complementarities was a significant predictor of failure, suggesting that our baseline distance measure can be interpreted as capturing proximity in productive capabilities. This result is supported by domestic manufacturing data.

This paper gives rise to a number of questions related to the dynamics of country-level shifts in the export baskets. What are the roles played by new and existing firms? Are different firms failing when attempting to export unrelated products, or are there patterns that suggest

that some exports are close to being internationally competitive but that individual firms still fail at their attempts? Does the relationship between distance and survival depend on which products are exported to which destination markets? Hopefully, the results from this paper will be useful in informing theory and in motivating the development of more micro-level studies about changes in specialisation patterns.

References

Albornoz, F., Pardo, H. F. C., Corcos, G., and Ornelas, E. (2012). Sequential exporting. *Journal of International Economics*, *88*(1), 17–31.

Balassa, B. (1965). Trade liberalization and revealed comparative advantage. *Manchester School of Economic and Social Studies*, *33*(2), 99–123.

Bernard, A. B., Jensen, J. B., Redding, S. J., and Schott, P. K. (2009). The margins of us trade. *American Economic Review*, *99*(2), 487–93.
URL `http://ideas.repec.org/a/aea/aecrev/v99y2009i2p487-93.html`

Bernard, A. B., Redding, S. J., and Schott, P. K. (2010). Multiple-product firms and product switching. *The American Economic Review*, *100*(1), 70–97.

Bernard, A. B., Redding, S. J., and Schott, P. K. (2011). Multiproduct firms and trade liberalization. *The Quarterly Journal of Economics*, *126*(3), 1271–1318.

Besedeš, T., and Prusa, T. J. (2006a). Ins, outs, and the duration of trade. *Canadian Journal of Economics/Revue canadienne d'économique*, *39*(1), 266–295.

Besedeš, T., and Prusa, T. J. (2006b). Product differentiation and duration of us import trade. *Journal of International Economics*, *70*(2), 339–358.

Boehm, J., Fornaro, L., and Dhingra, S. (2016). Swimming upstream: input-output linkages and the direction of product adoption.

Brenton, P., Pierola, M. D., and von Uexkull, E. (2009). The life and death of trade flows: understanding the survival rates of developing-country exporters. *Breaking into Markets: Emerging Lessons for Export Diversification*, (pp. 127–44).

Cameron, A. C., and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Córcoles, D., Díaz-Mora, C., and Gandoy, R. (2014). Product sophistication: A tie that binds partners in international trade. *Economic Modelling*, *44*, S33–S41.

Eckel, C., Iacovone, L., Javorcik, B., and Neary, J. P. (2015). Multi-product firms at home and away: Cost-versus quality-based competence. *Journal of International Economics*, *95*(2), 216–232.

Eckel, C., Iacovone, L., Javorcik, B., and Neary, J. P. (2016). Testing the core-competency model of multi-product exporters. *Review of International Economics*, *24*(4), 699–716.

Eckel, C., and Neary, J. P. (2010). Multi-product firms and flexible manufacturing in the global economy. *The Review of Economic Studies*, *77*(1), 188–217.

Engelsman, E. C., and van Raan, A. F. (1994). A patent-based cartography of technology. *Research Policy*, *23*(1), 1–26.

Evenett, S., and Venables, A. (2002). Export growth in developing countries: market entry and bilateral trade flows. Tech. rep., Mimeo, World Trade Institute, Bern.

Fan, J. P., and Lang, L. H. (2000). The measurement of relatedness: An application to corporate diversification. *The Journal of Business*, *73*(4), 629–660.

Farjoun, M. (1994). Beyond industry boundaries: Human expertise, diversification and resource-related industry groups. *Organization science*, *5*(2), 185–199.

Fernandes, A. M., and Paunov, C. (2015). The risks of innovation: Are innovating firms less likely to die? *Review of Economics and Statistics*, *97*(3), 638–653.

Fontagné, L., Secchi, A., and Tomasi, C. (2016). The fickle fringe and the stable core: Exporters' product mix across markets.

Fu, D., and Wu, Y. (2014). Export survival pattern and its determinants: an empirical study of chinese manufacturing firms. *Asian-Pacific Economic Literature*, *28*(1), 161–177.

Gorg, H., Kneller, R., and Murakosy, B. (2012). What makes a successful export? evidence from firm-product level data. *Canadian Journal of Economics*, *45*(4), 1332–1368.

Hausmann, R., and Klinger, B. (2007). The structure of the product space and the evolution of comparative advantage. *Harvard University Center for International Development working paper*, (146).

Hess, W., and Persson, M. (2012). The duration of trade revisited. *Empirical Economics*, *43*(3), 1083–1107.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, *317*(5837), 482–487.

Hummels, D., and Klenow, P. J. (2005). The variety and quality of a nation's exports. *American Economic Review*, *95*(3), 704–723.
URL `http://ideas.repec.org/a/aea/aecrev/v95y2005i3p704-723.html`

Iacovone, L., and Javorcik, B. S. (2008). Multi-product exporters: Diversification and micro-level dynamics. *World Bank Policy Research Working Paper Series, Vol.*

Iacovone, L., and Javorcik, B. S. (2010). Multi-product exporters: Product churning, uncertainty and export discoveries. *The Economic Journal*, *120*(544), 481–499.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford bulletin of economics and statistics*, *57*(1), 129–136.

Jenkins, S. P. (2008). Survival analysis with stata online course, available at http://www.iser.essex.ac.uk/survival-analysis. `http://www.iser.essex.ac.uk/survival-analysis`. Accessed: 10-10-2015.

Lejour, A. (2013). The duration of dutch export relations: decomposing firm, country and product characteristics. *De Economist*, *163*(2), 155–176.

Levinsohn, J., and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, *70*(2), 317–341.

Mayer, T., Melitz, M., and Ottaviano, G. (2014). Market size, competition, and the product mix of exporters. *American Economic Review*, *104*(2), 495–536.
URL `http://EconPapers.repec.org/RePEc:aea:aecrev:v:104:y:2014:i:2:p:495-536`

Navarro, L. (2012). Plant level evidence on product mix changes in chilean manufacturing. *The Journal of International Trade & Economic Development*, *21*(2), 165–195.

Neffke, F., and Svensson Henning, M. (2008). Revealed relatedness: Mapping industry space. *Papers in Evolutionary Economic Geography*, *8*, 19.

Nelson, R. R., and Winter, S. G. (1982). An evolutionary theory of economic change. *Cambridge: Belknap*.

Nicoletti, C., and Rondinelli, C. (2010). The (mis) specification of discrete duration models with unobserved heterogeneity: a monte carlo study. *Journal of Econometrics*, *159*(1), 1–13.

Prahalad, C., and Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review*, *May-June 1990*.

Rauch, J. E., and Watson, J. (2003). Starting small in an unfamiliar environment. *International Journal of Industrial Organization*, *21*(7), 1021–1042.

Sueyoshi, G. T. (1995). A class of binary response models for grouped duration data. *Journal of Applied Econometrics*, *10*(4), 411–432.

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior & Organization*, *23*(1), 1–30.

Wagner, R., and Zahler, A. (2015). New exports from emerging markets: do followers benefit from pioneers? *Journal of Development Economics*, *114*, 203–223.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Appendices (NOT FOR PUBLICATION)

A 'Jumps' and distance: some examples

This table is only meant to give an idea of (i) what types of 'jumps' are being studied, and (ii) the meaning of the distance measure. The first column shows the product labels[40] for the goods exported by a firm in the year $t-1$, and the second column the label of a product added by the firm in the year t. The third column shows the pairwise distance between the two products $(1-\phi_{i,j})$.

[40] These examples use only HS codes that were not modified by the homologation procedure, so that their labels correspond to the original HS classification. 'nes' stands for *not elsewhere specified*.

Table A1: Examples of 'jumps'. Distance as defined in Equation (1)

Products exported by a firm in the year $t-1$	**Product added by the firm in the year t**	**Pairwise distance**
Tomato ketchup and other tomato sauces	Tomatoes, preserved otherwise than by vinegar	0.73
Other jams, fruit jellies, marmalades, etc, ...		0.84
Articles of glass nes		0.94
Other grape must, nes	Wine (not sparkling); grape must with by alcohol	0.53
Wine (not sparkling); grape must with alcohol		0.56
Virgin olive oil and fractions		0.68
Other fixed vegetable fats and fractions, nes		0.72
Other fermented beverages (for example, cider,...		0.75
Casks, barrets, vats, tubs, etc, and parts thereof		0.83
Stoppers, lids, caps and other closures of plastic		0.86
Sacks and bags (incl. cones) of polymers of ethanol		0.89
Textile articles for technical uses, nes		0.89
Men's or boys' trousers, breeches, etc, of cotton	Women's or girls' trousers, breeches, etc.	0.36
Jerseys, pullovers, etc, of cotton, knitted...		0.43
Bables' garments and clothing accessories of cotton		0.60
Other blankets and travelling rugs, nes		0.68
Shampoos	Beauty, make-up, skin-care (incl. suntan), nes.	0.50
Animal products, nes; dead animals of Chapter 1		0.85
Sweetened milk and cream (excl. in solid form)		1.00

Table B1: (log) Sales of new manufactured product (first year) and measures of distance.

Dep.var.: (log) sales	I	II	III	IV
Distance from the core	-0.878***	0.0142	-0.705***	-0.419*
	(0.002)	(0.947)	(0.009)	(0.078)
Weighted dist. from non-core	-1.315***	-0.713***	-1.287***	-2.111***
	(0.000)	(0.001)	(0.000)	(0.000)
Share of the firm's exports		4.645***		
		(0.000)		
Share of the country's exports			2.745***	
			(0.000)	
Ranking dummies				✓
N	5787	5787	5787	5787
R2	0.755	0.867	0.769	0.831

Notes. OLS regressions using Chilean manufacturing data (1995-2006). Each observation is an addition of a new 5-digit CPC code to a firm's production basket. The dependent variable is the log of sales of the new product during its first year. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. Weighted dist. from non-core is the weighted distance between the new product and all products exported in t-1 minus the distance from the core. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level.

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Distance, sales and survival for domestic sales

The theoretical predictions relating distance to revenues and survival hold for both domestic sales and exports.

Using the domestic manufacturing sales data described in Section 3.1, we replicate the analysis from Section 3.3. Tables B1 an B2 present the results for regressing sales and survival on the distance from the core measure,[41] the measure of weighted distance to the rest of the basket and the different variables that have been used as proxies for distance as controls. All regressions include firm, five-digit CPC code and year of the 'jump' dummies. The measure of distance from the core is not significant for sales when using the share of the firm's sales as a proxy for distance, but in all other cases, both the measures of distance from the core and the weighted distance from the rest of the basket are significant determinants of sales and survival.

This means that our measure of distance is also consistent with the predictions for domestic manufacturing, and that survival of new products in the domestic product mix also depends on our proposed distance measures. As this is based on manufacturing data, these results support the idea that the distance measure captures similarities in terms of the (multiple) competences that are required to produce different goods.

[41] Domestic manufacturing products are classified using 5-digit CPC codes. To obtain the distance measure for CPC codes, the 6-digit HS BACI world trade data is converted into 5-digit CPC codes, and pairwise distances are then calculated as described in the main text. We cannot use the concordance procedure, as there is no correspondence between our concorded HS codes and CPC codes; however, the risk of misclassification is reduced when the data is aggregated into broader categories.

Table B2: Survival of new product (years) and measures of distance.

Dep.var.: years survived	I	II	III	IV	V
Distance from the core	-1.727***	-1.551***	-1.464***	-1.670***	-1.634***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Weighted dist. from non-core	-1.589***	-1.324***	-1.411***	-1.580***	-1.712***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(log) Initial value		0.201***			
		(0.000)			
Share of the firm's exports			1.373***		
			(0.000)		
Share of the country's exports				0.917***	
				(0.000)	
Ranking dummies					✓
N	5787	5787	5787	5787	5787
R2	0.609	0.619	0.619	0.611	0.612

Notes. OLS regressions using Chilean manufacturing data (1995-2006). Each observation is an addition of a new 5-digit CPC code to a firm's production basket. The dependent variable is the number of years that the new spell survives after its first appearance. Distance from the core is the distance between the new product and the most important product for the firm (in value) in t-1. Weighted dist. from non-core is the weighted distance between the new product and all products exported in t-1 minus the distance from the core. (log) initial value is the (log) export value of the new product. Share of the firm's exports is the % in total firm exported value of the newly exported product. Share of the country's exports is the weight of the new product exported by the firm in the country's exported value of the product. The ranking dummies ranks by value of exports the *new* products added each year by each firm.

All regressions include firm, product code and year dummies. Standard errors clustered at the firm level.
p-values in parentheses. * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01.

C Survival analysis

In survival or duration analysis, the survival function captures the unconditional probability that an 'individual' (a new firm-product export flow) survives up to t years (t is not calendar time, but the number of years that the export flow survives). Formally it is defined as $S(t) = Pr\{T > t\}$, where T is the moment where the 'event' (the 'death' of the new export) occurs. The simplest survival analysis method is the nonparametric Kaplan-Meier estimator of the survival function, which accounts for censoring, but cannot incorporate the effect of covariates.

Survival analysis usually focuses on the *hazard function*, which defines the probability of 'dying' in a certain moment, conditional on having survived up to that moment. Accounting for covariates, the benchmark model in survival analysis is Cox's proportional hazards model. For this model, the hazard $\theta_{i,t}$ of individual i at time t is defined as:[42]

$$\theta_i(t, \mathbf{x}_{i,t}) = \lim_{dt \to 0} \frac{Pr\{tT < t + dt | T \geq t\}}{dt} = \theta_0(t) exp\left(\mathbf{x}'_{i,t}\beta\right) \tag{5}$$

where $\theta_0(t)$ is called the baseline hazard and $\mathbf{x}_{i,t}$ is a vector of observed covariates. The hazards are said to be proportional because the semielasticity of the covariates on the hazard rates are constant across t.[43] In other words, changes in covariates have a multiplicative effect over the baseline hazard rate, and these effects are the same regardless of how long the product

[42] The hazard function can also be expressed in terms of the survival function, $\theta(t, \mathbf{x}) = -S'(t)/S(t)$.
[43] In equation (5), $\partial ln(\theta)/\partial x_{tk} = \beta_k$.

has already 'survived'. For instance, proportional hazards mean that differences in the initial value of export flows have the same proportional effects on the risk of the flow stopping in the second or the tenth year after it started. However, one would expect the effect of the initial value to be relatively higher during the first years.

Cox's model is popular because β can be estimated using a partial likelihood approach, without any assumption about the shape of the baseline hazard (it is a semi-parametric model).

But this model has three important shortcomings, as discussed in detail by Hess and Persson (2012): (1) estimates of the coefficients and their standard errors are biased under many tied durations, which is the norm in trade flow duration data;[44] (2) it is very difficult to account for individual unobserved heterogeneity and (3) the assumption of proportional hazards is often incorrect (either because of inherent non-proportionality or because of unobserved heterogeneity which is not accounted for).

It is possible to use a discrete-time equivalent of the Cox model, the complementary log-log model (or *cloglog*) with random effects, to account for (1) and (2), but to account for these and simultaneously relax (3), which is tested and rejected by Brenton et al. (2009) and Hess and Persson (2012) in a trade flows context, other models must be used.

Hess and Persson (2012) argue that the best way to deal with the three issues simultaneously is to use panel binary response models, where the dependent variable is whether an individual switches state or not. This approach is followed by several papers on the duration of trade flows, such as Fernandes and Paunov (2015), Fu and Wu (2014) and Lejour (2013).

Jenkins (1995) and Sueyoshi (1995) show that defining a dummy variable $y_{i,k}$, where i indexes the trade spells and k the possible time intervals, equal to zero on periods where a trade spell i is 'at risk', and equal to one on the period where observation i changes state (k_i), the log-likelihood of the survival function for discrete duration data can be written as:

$$ln(L) = \sum_{i=1}^{n}\sum_{k=1}^{k_i} \left[y_{i,k} ln\left(h_{i,k}\right) + \left(1 - y_{i,k}\right) ln\left(1 - h_{i,k}\right)\right] \tag{6}$$

That is, the log-likelihood of the survival model is isomorphic to that of panel binary response models, where the hazard $h_{i,k}$ takes the place of the cumulative density function of the unobserved in a binary model.

The discrete hazard function for the time interval k, i.e. the probability that the spell i, characterised by $\mathbf{X}_i$, ends during the time interval k, conditional on having survived up to the beginning of k, is:

$$h_{i,k} = Pr\left[T_i < t_{k+1} | T_i \geq t_k, \mathbf{x}_{i,k}\right] \tag{7}$$

where T_i is the time at which the spell ends, t_k the lower limit of each time interval k, and $\mathbf{x}_{i,k}$ a set of possibly time-varying controls that can be stacked into $\mathbf{X}_i$.

To maximise the likelihood in equation (6) it is necessary to define a functional form for the discrete-time hazard h_{it}. The most commonly used functional forms are the cumulative density

[44] The Cox model is for continuous time, where the probability of identical durations is zero. With discrete yearly data all observations will have a duration equal to many others.

functions of the type-I extreme value, logistic and normal distributions, which lead to the cloglog, logit and probit models respectively. The covariates enter these distribution functions as a linear index, so that for instance for the logit model the hazard function takes the following form:

$$h_{i,k} = F\left(\mathbf{x}'_{i,k}\beta+\gamma_k\right) = \frac{\exp(\mathbf{x}'_{i,k}\beta+\gamma_k)}{1+\exp(\mathbf{x}'_{i,k}\beta+\gamma_k)} \tag{8}$$

where the set of γ_k dummies represent the nonparametric baseline hazard function (equivalent to the $\theta_0(t)$ in equation (5) for continuous time). In other words, they represent the base risk of dying each period, regardless of the value of the covariates, and it is estimated without imposing any functional form on it. With the cloglog model the effects of covariates on the hazard are approximately proportional and constant across the life of the trade spell. While in theory this proportionality of hazards is relaxed both by the probit and logit models, Sueyoshi (1995) shows that the logit hazards are very close to being proportional, and only for the probit model the marginal effects of covariates on the hazard have significant heterogeneity across the time periods k, effectively producing non-proportional hazards.

For consistent estimation of the parameters, the likelihood must be correctly specified. This requires that spells are conditionally independent and that censoring is exogenous (in the sense of providing no information about survival time beyond what the covariates do). To achieve the former it is necessary to control for the possible dependencies, and for the latter, to include a set of dummies for the starting year of each spell (this accounts for a problem induced by the censoring occurring at a fixed calendar date).

When time-varying covariates are included, they must be strictly exogenous in the sense that the hazard in period k depends only on the covariates included in $\mathbf{x}_{i,k}$ (which can include lagged covariate values, see Wooldridge, 2010).

There is another way in which the likelihood is likely to be misspecified: the distributional assumption on the hazard implies that it is completely determined by the observed covariates, without room for an error term. The models described above can relax this by including random effects.

The random effects enter as a multiplicative factor on the hazard function (with mean one). In proportional hazard models, this can also be expressed as a linear index of the form $\mathbf{x}'_{i,k}\beta + \gamma_k + \varepsilon$, where ε (which has mean zero) can be interpreted as summarising the impact of omitted variables on the hazard (Jenkins, 2008). Usually a parametric distribution is specified for ε, which can then be integrated out of the likelihood for estimation (Jenkins, 1995). Nicoletti and Rondinelli (2010) show with simulations that parameter estimates are robust to mistakenly using a normal distribution when the true distribution is gamma or discrete. As Hess and Persson (2012) point out, this finding is supported by empirical studies that show that the choice of distribution is not important, as long as the baseline hazard is modelled nonparametrically. In the regressions reported below, when random effects are included they are assumed to come from a Normal distribution.[45]

[45] Random effects must be independent from (not only uncorrelated to) the covariates. The only way to relax this and include fixed effects is when there are several spells observed for each unit of observation (Wooldridge, 2010). This could be attempted with re-entries, but sample size would drop dramatically.

D Marginal effects for nonlinear models

Table D1: d(ln[h])/d(Weighted distance) at different survival times for random effects models (with interaction)

	RE Clog-log	RE Logit	RE Probit
Weighted distance			
at survival=0	0.294***	0.317***	0.296***
	(0.000)	(0.000)	(0.000)
at survival=1	0.196***	0.177***	0.158***
	(0.000)	(0.000)	(0.000)
at survival=2	0.186***	0.163***	0.147***
	(0.000)	(0.000)	(0.000)
at survival=3	0.174***	0.147***	0.133***
	(0.000)	(0.000)	(0.000)
at survival=4	0.164***	0.136***	0.124***
	(0.000)	(0.000)	(0.000)
at survival=5	0.141***	0.112***	0.105***
	(0.000)	(0.000)	(0.000)
at survival=6	0.119***	0.0928***	0.0899***
	(0.000)	(0.000)	(0.000)
at survival=7	0.0872***	0.0706***	0.0720***
	(0.000)	(0.000)	(0.000)
at survival=8	0.0526**	0.0521***	0.0566***
	(0.032)	(0.000)	(0.000)
at survival=9	0.0560**	0.0548***	0.0593***
	(0.044)	(0.000)	(0.000)
at survival=10	0.0249	0.0381***	0.0440***
	(0.276)	(0.001)	(0.000)
at survival=11	0.0262	0.0409**	0.0474***
	(0.407)	(0.014)	(0.004)
at survival=12	0.00593	0.0254**	0.0311**
	(0.645)	(0.033)	(0.023)
at survival=13	0.000138	0.0132	0.0163
	(0.847)	(0.100)	(0.129)
N	28572	28572	28572

Notes. Semielasticity of weighted distance on the conditional probability of surviving one more year at different survival times for each random-effects models in Table 9 for a typical jump (all covariates are evaluated at their means, except for the new 4-digit sector, jump year and HS2 dummies, evaluated at their modes.).

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D2: d(ln[h])/d(Weighted distance) at different survival times for random effects models

	RE Clog-log	RE Logit	RE Probit
Weighted distance			
at survival=0	0.265***	0.277***	0.258***
	(0.000)	(0.000)	(0.000)
at survival=1	0.236***	0.231***	0.209***
	(0.000)	(0.000)	(0.000)
at survival=2	0.223***	0.210***	0.191***
	(0.000)	(0.000)	(0.000)
at survival=3	0.208***	0.188***	0.172***
	(0.000)	(0.000)	(0.000)
at survival=4	0.195***	0.172***	0.159***
	(0.000)	(0.000)	(0.000)
at survival=5	0.167***	0.140***	0.133***
	(0.000)	(0.000)	(0.000)
at survival=6	0.140***	0.114***	0.113***
	(0.000)	(0.000)	(0.000)
at survival=7	0.100***	0.0856***	0.0892***
	(0.000)	(0.000)	(0.000)
at survival=8	0.0588*	0.0622***	0.0691***
	(0.054)	(0.000)	(0.000)
at survival=9	0.0624*	0.0651***	0.0721***
	(0.070)	(0.000)	(0.000)
at survival=10	0.0262	0.0445***	0.0523***
	(0.329)	(0.001)	(0.000)
at survival=11	0.0271	0.0470**	0.0557***
	(0.455)	(0.016)	(0.006)
at survival=12	0.00550	0.0288**	0.0356**
	(0.677)	(0.037)	(0.032)
at survival=13	0.0000921	0.0146	0.0178
	(0.861)	(0.106)	(0.155)
N	28572	28572	28572

Notes. Semielasticity of weighted distance on the conditional probability of surviving one more year at different survival times for each random-effects model in Table 8 for a typical jump (all covariates are evaluated at their means, except for the new 4-digit sector, jump year and HS2 dummies, evaluated at their modes.).

p-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.