Firms in Product Space: Adoption, Growth and Competition

Luca Macedoni[a], John Morrow[b], Vladimir Tyazhelnikov[c]

25 January 2024

Preliminary Draft

Abstract

Which products are potentially produced together? When demand for a product increases, which firms will supply it? Using multi-product production patterns within and across firms, we recover a continuous cost based distance between firms and *unproduced* products. Higher product distance implies decreasing adoption frequency. When export demand induces domestic product adoption, closer firms provide this supply. Potential costs imply measures of Revenue and Competition Potential. These predict firm sales growth, scope growth and core focus. If all firms produced all products linked by co-production, consumer welfare would increase by 10-30%.

JEL Codes: F1, D2, L1, L23, L25.
Keywords: Multi-product firms, firm capabilities, product classification, product space, growth paths.

Acknowledgments. We thank Nick de Roos for early suggestions and Axel Roepstorff for excellent research assistance. We are grateful for comments from Catherine Thomas, Kirill Borusyak, Andreas Hafler, Wanyu Chung, Magnus Tolum Buus, Guzman Ourens, Luca Lodi, and participants at the 1st SICKCL, Economics of Global Interactions, KCL Seminars, ETSG, LSE Trade Workshop, Kent Macro Workshop, Australasian Trade Workshop, WEAI, Nottingham, Zurich ETH, Paris Trade & Macro Workshop, Asia-Pacific Trade Seminar, PSE/Sciences Po, InsTED, Lille, Organisational Economics Workshop, UC Louvain, Tilburg GIFT Conference, BCAM Macro, CEP Retreat, ESCOE, ITSG, Workshop on the Impact of Global Trade Shocks, Firms/Workers/Inst Linz, RES, SMU, INSE at Peking University, Universitat Pompeu Fabra, HKS Growth Lab, NYU Trade Day, CUHK Trade Workshop, Geneva Grad Institute, EUI Trade and Development Workshop, QMUL, and the Competition and Markets Authority.

[a]Aarhus University. Financial support from the Carlsberg Foundation is gratefully acknowledged.
[b]Kings College London, CEP/LSE, and CEPR. Corresponding Author: john.morrow@kcl.ac.uk.
[c]The University of Sydney.

1 Introduction

Which products are potentially produced together? When demand for a product increases, can we predict which firms will supply it? This is important for policy as knowing ***which*** products a firm is likely to adopt can help target policy when key supplies are desired, such as "green" products or those important during an emergency. A key obstacle is that until a firm produces a product, its potential costs are unknown. We introduce a novel approach to predict such costs based on observed co-production patterns within and across firms. We represent these costs as the distance from firms to products in a multi-dimensional product space.

Mapping the product space is the problem opposite to triangulation: finding the position of cell phone users (products) by their distance from cell phone towers (firms).[1] In our application, we derive firm-product costs either from inverting demand or from production based methods akin to De Loecker et al. (2016). We map these costs to distances, but do not know any firm distances to unproduced products. Using observed firm-product distances, we back out pairwise distances between products to infer unknown firm-product distances. These distances then map to costs, providing potential costs for all firm-product pairs.[2] Constructing this high-dimensional ***product space*** is a new approach that allows for novel counterfactuals based on standard theory and empirical strategies from the heterogeneous firm literature.

We employ rich Danish firm-level data to validate our approach and test two predictions about ***which*** firms produce ***which*** products. First, firm distance to a product predicts adoption in a highly granular way, even controlling for discrete distances embedded in the Harmonized System.[3] Second, we validate the prediction of which firms will supply products by instrumenting product level demand with export demand, finding

[1]In computer science, the analysis of self-positioning networks leads to a similar problem. For example, indoor positioning algorithms use the information on signal strength from multiple Wi-Fi routers with a priori unknown locations to identify the location of a cell phone user with a weak GPS signal (see for example Wu et al. (2005) and Hossain and Soh (2015)). Note, however, that these problems are typically solved in a 2- or 3-dimensional space, while we focus on a high-dimensional environment.

[2]Unlike countries, the production patterns of firms are sparse relative to the number of products. A contribution of our paper is to develop an algorithm that relies on co-production *within and across* firms, so as to generate a complete set of product-to-product distances even though the majority of product pairs is never co-produced within a firm.

[3]In contrast with most studies that look *within* the firm and rely on product classifications as measures of closeness, we arrive at them empirically rather than from statistical hierarchies which embody a mixture of rationales. See the examples of Jacobs and ONeill (2003) and Grant (2023). Table A.1 highlights differences across the HS, SIC and NAICS. The organizing principles of classification systems vary: "NAICS differs significantly from the SICs because it is based on a single organizing principle, contrary to the SICs where entities are sometimes grouped according to production-oriented principles and sometimes grouped according to demand-based principles. NAICS is based on a production-oriented or supply based conceptual framework [...] very similar production processes are grouped." (Girard and Trau, 2004).

that a positive demand shock induces the closest firms to introduce a product.

Moreover, the product space reveals opportunities and threats. Firms close to unproduced products may have ample space for expansion, while those far from potential products may have low growth potential. The trade and growth literature has demonstrated that countries with better market access through better trade channels grow faster (Redding and Venables, 2004); and while the international trade literature has ample measurements of distance to markets, there is no comparable concept of firm distances to unproduced products. We introduce a Revenue Potential (RP) index of how much a firm could increase its revenue by producing all accessible products in the product space. RP predicts both sales and scope growth towards a more diverse product portfolio.

We apply similar logic to potential competition in each firm's existing product portfolio and calculate a Competition Potential (CP) index. While it might seem that a firm is positioned to expand rapidly into nearby products, it is possible that it is surrounded by a large number of potential entrants, restricting scope growth and potentially sales growth. The CP index calculates how a firm's revenue would fall if all potential rivals choose to compete with a firm in all its existing markets. In fact, high CP restricts scope growth and portfolio diversity, but sales growth is unchanged.

Finally, counterfactual costs allow measurement of potential gains for consumers by predicting the impact of firm entry into unproduced varieties on the price index. We define Entry Potential (EP) to capture the upper bound for such gains. The predicted gains are heterogeneous by sector, ranging from 10-30%.

This paper proceeds as follows. This Section continues with a literature review, and the next motivates and theoretically constructs the ***product space***. Section 3 then details the data and construction of the space with summary statistics of its properties. Section 4 estimates how the product space can predict product adoption. Section 5 models Revenue Potential and Competition Potential and estimates their impact on sales growth, scope growth and core focus and quantifies potential changes in the cost of living from production of all products by all firms with Entry Potential. Section 6 concludes.

Literature Review

Underlying much of trade theory is production theory, whether across neoclassical models or the leap to New Trade theory with monopolistic competition and increasing returns to scale. 'New' New Trade theory augmented production structures with heterogeneous firms and sales at the firm level, and multi-product firm models have become even more granular with heterogeneous sales activity within the firm. All of these models study firm production in isolation. However, the distribution of activities firms engage in together is not random, suggesting a richer production setting of interconnections across activities. In the language of comparative advantage, opportunity costs across activities are not distributed randomly, whether or not firms engage in the activities themselves. Our approach is to uncover such a latent map of potential costs to provide a rich set of counterfactuals about the growth potential and competitiveness of firms, as well as new measures of potential consumer surplus. By doing so, we connect a few well developed strands of literature across multi-product firms, ideas of co-production relatedness between products at the firm and country export basket levels, and classification systems.

There has been an explosion of research on multi-product firms, especially in the context of international trade.[4] For the typical model of this literature, a firm is a collection of products, which may be linked by supply or demand linkages.[5] A product is defined as a variety produced by one firm and is characterized by marginal costs of production and/or demand shifters (Eckel and Neary, 2010; Bernard et al., 2011). However, which particular product a firm produces is essentially neglected: products differ in their sales and this difference is driven by differences in costs or demand. Whether a firm is producing milk and cheese or milk and silk is irrelevant. By recovering distances in a product space based on co-production costs, ***which*** products a firm produces relative to ***all other rms*** matters for sales and scope growth.

The literature on co-production (Bernard et al., 2010; Goldberg et al., 2010) shows that some pairs of products are often produced together, while others are almost never produced by the same firm. In this approach, when a firm expands its product range, it will likely choose products that are often co-produced within other firms, conditional on their current product mix, for a wide variety of possible explanations for linkages across inputs and outputs of firms (Boehm et al., 2022; Jakel et al., 2023). We build on such co-production concepts by using observed costs when products are co-produced to quantify the cost side role of co-production.

The concept of a product space has been popularized by the groundbreaking and continuing work of authors such as Hidalgo et al. (2007) and Hausmann et al. (2007) and has recently been applied to the case of green products (Mealy and Teytelboym, 2022). One key finding from these studies is that differences in income growth rates across countries can be explained by their proximity to various products.[6] Our paper

[4]For details, see the recent review by Irlacher (2022).

[5]Supply linkages include flexible manufacturing, economies and diseconomies of scope, and the presence of core and non-core products (Eckel and Neary, 2010; Nocke and Yeaple, 2014; Mayer et al., 2014; Eckel et al., 2015; Arkolakis et al., 2021; Macedoni and Xu, 2022). Demand linkages mainly include cannibalization effects and demand complementarities (Feenstra and Ma, 2007; Eckel and Neary, 2010; Dhingra, 2013; Bernard et al., 2018; Flach and Irlacher, 2018; Macedoni, 2022).

[6]Hidalgo et al. (2007) show that countries tend to develop an RCA in goods closer to the ones they are currently specialized in. As a result, countries that are closer to non-RCA goods upgrade their exports quicker. This mirrors our findings that firms tend to begin producing products that are closer to them and that higher Revenue Potential leads to faster growth.

diverges from this literature in two significant aspects. First, we focus on firms rather than countries, providing a micro-foundation for the country-level position in the product space. Second, unlike in Hidalgo et al. (2007) where the distance between products is determined by the co-presence of Revealed Comparative Advantage (RCA), we base our measure on marginal costs. While our algorithm can construct a product space and firm space using RCA or other metrics, using marginal costs allows us to interpret distances in terms of actual or potential marginal costs, and to construct microfounded measures of market potential and competition. Furthermore, our emphasis on co-production within and across firms represents a significant methodological advancement, addressing a key limitation in export baskets based approaches when applied to firms due to the sparsity of actually produced products at the firm rather than country level. Finally, the micro foundation we propose provides a supply curve of successive firm production of any particular variety, which is more granular than binary measures, such as whether a variety exhibits RCA.

Our focus on firms in a product space links our paper to an expanding body of research that uses various measures of distance or similarity *between firms* for different applications. For example, in the R&D literature, the location of firms is crucial for evaluating spillovers across firms, which depend on the proximity between their technologies and products. A common method in this field involves calculating the overlap between firms' technology classes, as indicated by patents, and the overlap between their product sales (Jaffe, 1986; Bloom et al., 2013).[7] Additionally, recent work by Pellegrino (2019) develops a model in which a firm's demand and market power depend on the similarity of its products to those of oligopolistic competitors. The author uses the similarity between products developed by Hoberg and Phillips (2016) that we discuss below. Our paper differs from this literature by focusing solely on the distances between firms and products, as these represent either actual or potential marginal costs. Our analysis captures the impact of competitors on a firm through the metrics of Revenue Potential and Competition Potential, which are influenced by the proximity of the competitor to the firm's current or potential products.

A small but growing set of research has been creating new categorizations of firm activities and outputs, often using advances in text analysis to uncover new relationships. In an exciting strain of work, Hoberg and Phillips (2016) create firm locations from word vectors of SEC filings, in which firms have new relative locations each year based on cosine similarity measures which generally better explain profitability and growth than SIC or NAICS classifications. Looking at the shocks of post-9/11 military spending increases and

[7]Escolar et al. (2023) provides a summary of the approaches used in the literature for locating firms using patents data.

the millennium tech crash suggests firms move to areas of common high demand or reduce similarity to differentiate after a negative demand shock. In contrast to this methodology, we use standard product classification codes in common administrative data which allows us to bring to bear well established techniques in the literature (such as widely accepted IV strategies) to a large potential range of datasets. Kogan et al. (2021) categorize worker's technology exposure from patent documents and the Dictionary of Occupational Titles, finding exposure displaces both high and low skill workers. Bishop et al. (2022) match UK business website data to publicly reported SIC codes, using text analysis to show that four digit SIC codes mask considerable heterogeneity in firm activities. In fact, governments endogenously change classifications to suit their objectives. Grant (2023) models and demonstrates that the US Harmonized Trading System is endogenous, with more product differentiation with higher trade flows or tariffs where mis-classification is more costly.

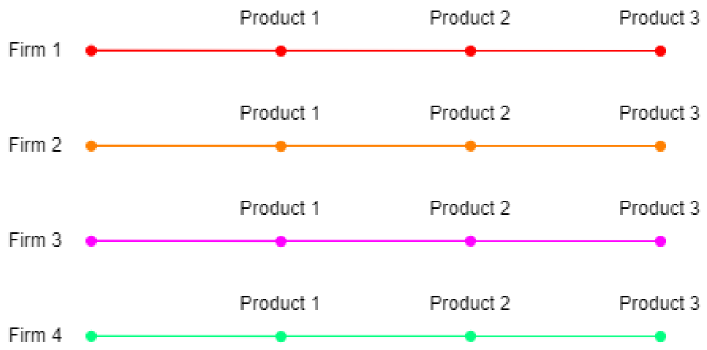
2 Firms in Product Space

This section models firms and products co-located in space to establish a concept of distance between products and between firms and products. Our product space is based on marginal costs and we assume these costs are positively related to the distances between firms and products. In this section, we will take marginal costs as given. Later in the paper we will estimate the marginal costs using two standard approaches from the literature: 1) we specify Constant Elasticity of Substitution (CES) demand and monopolistic competition in order to map the revenue shares of a multi-product firm to its marginal costs of production and 2) we obtain marginal costs from the markup estimation literature as in De Loecker et al. (2016). Moreover, we construct a discrete benchmark of distances from the Combined Nomenclature product hierarchy. The result is a continuous product classification that explains heterogeneous firm marginal costs across products in terms of each firm's location in relation to products in a high-dimensional product space.

2.1 An Example from the Multi-product Firm Literature

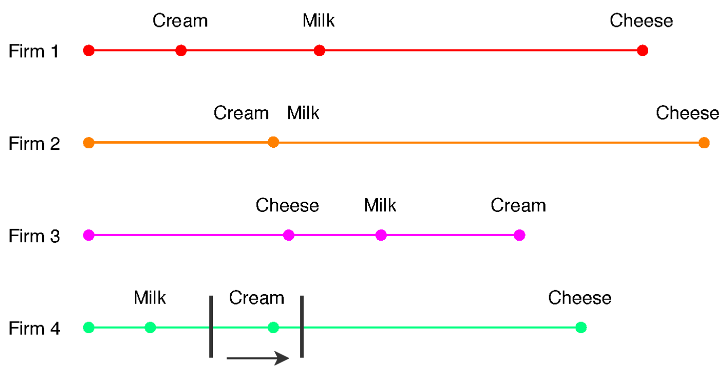
A well understood setting for co-occurring activities within firms are multi-product firms due to the incidental detailed data which provide evidence of differentiated activities. Such data has given rise to standard models of multi-product firms as typified by Eckel and Neary (2010) and several subsequent contributions. In this and similar settings, theory ranks products within a firm according to their marginal costs, which increase for products farther from the firm's core competence, but do not depend on *which* products are co-produced. This kind of setting can be visualized as in Figure 1, where the further a product is from a firm, the higher the cost.

Figure 1: Multi-product Costs



As an example of our framework, we keep the idea of a ranking of products based on their marginal cost but assume the existence of a product space in which the proximity of products and the pattern of production within the firm can alter these regular costs differences. We consider three products: Milk, Cream and Cheese and four firms which, due to different latent capabilities, have different costs for each product which occur at irregular intervals as in Figure 2.

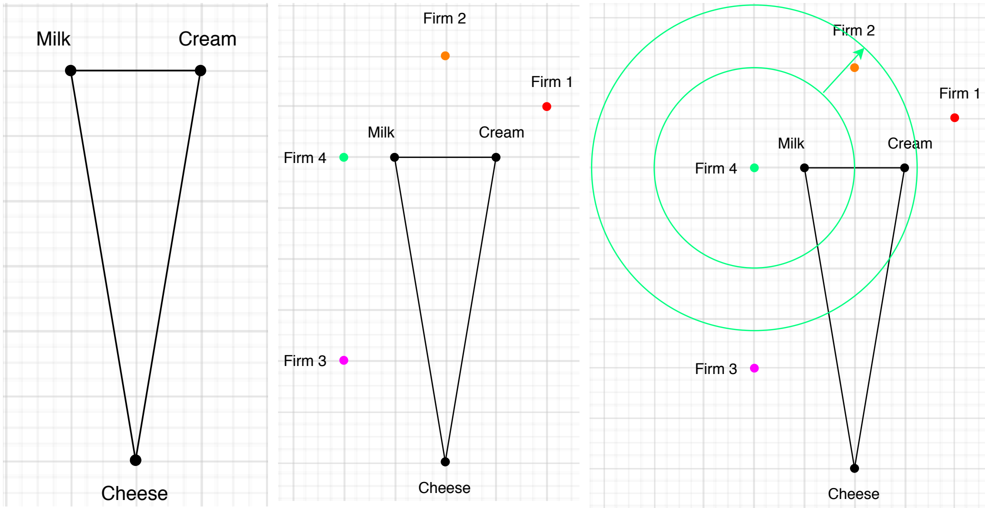
Figure 2: Firm and Product Costs



One way to rationalize this pattern is to locate the three products and the four firms in a two-dimensional space. As shown in Figure 3a, milk and cream have more similar production technologies, while cheese production is technologically distinct from either. This is represented in our framework with a small distance between milk and cream and a large distance of these away from cheese.

Figure 3: Technological Nexus

(a) Product Nexus (b) Firm Locations (c) Product Adoption



Every firm is characterized by its position relative to products as in Figure 3b. Here Firm 1 is relatively good at cream production, less good at milk production and the least good at cheese production. Firm 2 is equally good in the production of milk and cream, but not as good at cheese production. Firm 3 is good in the production of cheese, almost as good in the production of milk, and the worst in the production of cream. These relationships are represented in a more conventional way in Figure 2.

In this example, we make two assumptions, which, however, are not imposed in our empirical estimation of the product space. First, we assume identical aggregate demand for each variety each period. Second, the fixed cost of adding any product is constant across varieties. These assumptions allow us to represent a firm's adoption threshold as a single circle. This captures that when choosing between similar markets, a firm would choose to adopt a product with a lower marginal cost. Considerations of market size and competition intensity may shape this decision too. In a general setting that we rely on, a firm has a separate threshold for each good it can produce; these thresholds can be graphically represented as a series of concentric circles with bigger circles corresponding to larger and less competitive markets. In the empirical section we will either address this with time specific fixed effects or explicitly by constructing corresponding price indexes.

2.2 An Algorithm to Construct the Product Space

Here we present the algorithm we use to construct the product space and locate firms in it. Given a set of observed distances between firms and the products they produce, our goal is to estimate the distance of firms to the products they do not produce. To achieve this, we first estimate the distance between any two product pairs. Second, we locate products in the product space and, finally, we locate firms in the space. Our approach accounts for relative product costs of a firm by locating it at a different distance from each product.

Let ν denote a differentiated product and ω a firm. As in the conventional approach in heterogeneous firm literature (Melitz, 2003), each firm is characterized by a productivity φ_ω and can produce a unique variety of N differentiated products. For instance, a differentiated product may be cheese and a firm can make its own variety of cheese. Firm capabilities to produce products are characterized by relative costs $c_{\omega,\nu}$ for $\nu = 1, ..., N$. We observe the marginal costs of production for each product ν by firm ω, denoted $\frac{c_{\omega,\nu}}{\varphi_\omega}$. Normalizing absolute costs by φ_ω avoids a multiplicity of firms' locations corresponding to the same relative costs, and in the relevant estimates below, we will control for firm-time fixed effects.

Conditional on observing positive revenues $R_{\omega,\nu} > 0$, in the absence of error we would observe the set $C_\omega \equiv \left\{ \frac{c_{\omega,\nu}}{\varphi_\omega} : \quad R_{\omega,\nu} > 0 \right\}$. Given observed marginal costs, our goal is to map counterfactual relative costs for unsold products: $\Gamma_\omega \equiv \left\{ \frac{c_{\omega,\nu}}{\varphi_\omega} : \quad R_{\omega,\nu} = 0 \right\}$. The mapping $C_\omega \to \Gamma_\omega$ could contain arbitrarily rich economies of scope from co-production. For instance, the marginal costs of adding cheese to the product mix could differ depending on whether the firm is already making milk or cream. The set of possible combinations can be computationally intractable: considering the extensive margin alone, with around 1000 products, there are 2^{1000} possible permutations of production decisions just for a single firm. To deal with this complexity, the mapping $C_\omega \to \Gamma_\omega$ is fixed by locations of products and firms in a high-dimensional space where unknown relative firm costs are fixed by distance.

While each firm makes a subset of products, the collection of all firm-product observations reveals information regarding the distance of products from one another, so long as there are chains of co-production observed between any two products. These chains could be direct, i.e., the same firms co-produces two products, or indirect, i.e., two firms make the same product and all other products they make are indirectly linked. We define each set of products connected this way over all years as a ***cluster***. In the remainder, we focus on the algorithm to estimate the space for a single cluster, with N products.

To facilitate further analysis, we represent each product and firm by a location in $N-1$-dimensional product space. In fact, in the example of Section 2.1, the product space with three products (milk, cream, and cheese) can be represented in a two-dimensional space. The relative location of a firm to products determines its cost structure. The location of a variety ν is $\ell_\nu \in \mathbb{R}^{N\ \ 1}$ and the location of a firm ω is $\ell_\omega \in \mathbb{R}^{N\ \ 1}$. The cost based distance between any pair of products or firms is Euclidean, for which we use norm notation $\|\ell_\nu - \ell_\omega\| = \left(\sum_{i=1}^{N} {}^{1} \left(\ell_\nu^i - \ell_\omega^i\right)^2\right)^{1/2}$.[8] Firm-product distances are assumed to be observed with multiplicative error $\varepsilon_{\omega,\nu}$ and given by $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu}$. In the empirical section we map multiple data choices and coordinate systems from co-production cost data to $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu}$, so for the moment, we take them as observed cost based distances. Once we have constructed the locations of products, we will fix firm locations.

2.2.1 Bounding Product-to-Product Distances

If all firms produced all products in no matter what quantity, it would be possible to write down a system of equations where the distances from each firm to each product are known. No population of firms is likely to exhibit this property and any sub-sample of all firms producing all products would suffer from selection issues. However, appealing to revealed production preference, we propose an alternative procedure based on the triangle inequality which allows us to evaluate upper and lower bounds of distances when firms produce two or more goods.

Intuitively, if two products are never produced together, it suggests they are unlikely to be produced together in the future. Conversely, if the two products are always produced together, it suggests that there is something specific about such a pair. We extend this logic and look not only at co-production, but also at the intensity of co-production within firms. To quantify this, we use our product space framework and apply the triangle inequality to each firm-product distance pair of co-produced products.

While an application of the triangle inequality to a multi-product firm do not tell us exactly how far apart any two products it produces are, it does provide upper and lower distance bounds as shown in Figure 4a. This can be seen from the triangle inequality. Given two products ν and ν^0 produced by a firm ω, one can think about a two-product firm as a dot with two concentric circles representing the potential locations of both products. It is clear then that the shortest possible distance between these products is equal to the difference of two radii and the longest to their sum as illustrated in Figure 4b.

In the language of the triangle inequality, co-production within firms implies that the distance between two products ν and ν^0 absent measurement error, $d_{\nu\nu^0}$, satisfies

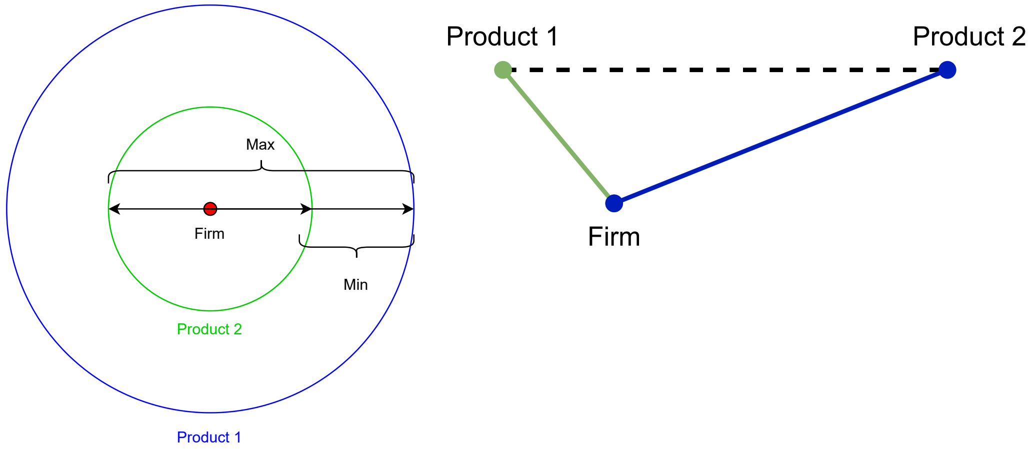
$$\left|\frac{c_{\omega,\nu}}{\varphi_\omega} - \frac{c_{\omega,\nu^0}}{\varphi_\omega}\right| \leq d_{\nu\nu^0} \leq \frac{c_{\omega,\nu}}{\varphi_\omega} + \frac{c_{\omega,\nu^0}}{\varphi_\omega}.$$

For each combination of a firm and a co-produced product pair, we construct maximum and minimum distances. Even one firm ω can give us narrow bounds if it is close to one

[8] This could be any norm, for instance the class of L^p norms could be chosen for goodness of fit.

Figure 4: Distances Between Directly Connected Products

(a) Max and Min Product Distance (b) Triangle Inequality Distance Bounds



product ν' (a small radius in Figure 4a) since for small $\frac{c_{\omega,\nu'}}{\varphi_\omega}$, the inequalities on both sides become tight.

Following this logic, we calculate the averages of the upper and lower product distance bounds across firms, $\overline{d}_{\nu,\nu'}$ and $\underline{d}_{\nu,\nu'}$ between goods ν and ν' as follows:

$$\overline{d}_{\nu,\nu'} \equiv \text{mean}_{\omega \in \Omega_{\nu,\nu'}} \left\{ \frac{c_{\omega,\nu}}{\varphi_\omega} \varepsilon_{\omega,\nu} + \frac{c_{\omega,\nu'}}{\varphi_\omega} \varepsilon_{\omega,\nu'} \right\},$$

$$\underline{d}_{\nu,\nu'} \equiv \text{mean}_{\omega \in \Omega_{\nu,\nu'}} \left\{ \left| \frac{c_{\omega,\nu}}{\varphi_\omega} \varepsilon_{\omega,\nu} - \frac{c_{\omega,\nu'}}{\varphi_\omega} \varepsilon_{\omega,\nu'} \right| \right\},$$

where $\Omega_{\nu,\nu'}$ is a set of firms co-producing goods ν and ν'. Within each cluster, we then define the distance $d(\nu,\nu')$ for co-produced products as the average of $\overline{d}_{\nu,\nu'}$ and $\underline{d}_{\nu,\nu'}$:[9]

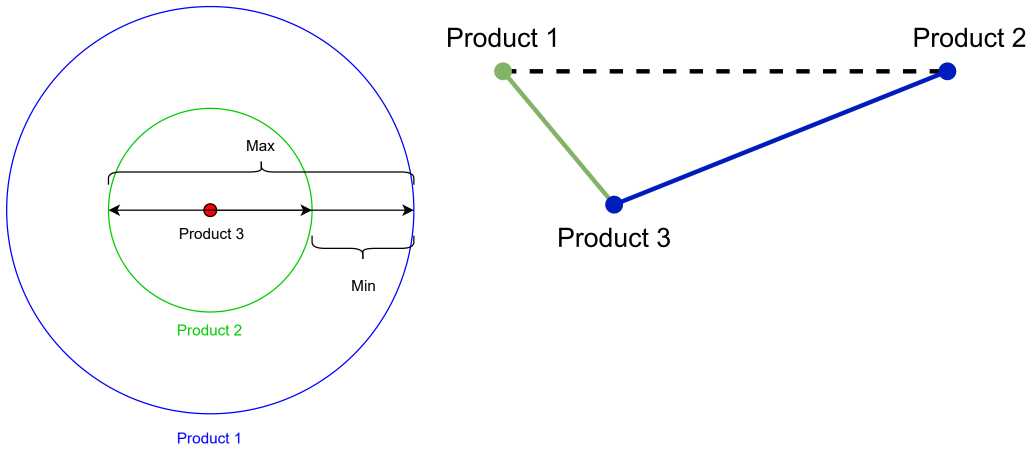
$$d_{\nu,\nu'} \equiv \left(\overline{d}_{\nu,\nu'} + \underline{d}_{\nu,\nu'}\right)/2.$$

Although application of the triangle inequality allows us to find the bounds for distances between any pair of co-produced products, for our approach to work, we need distances between ***all*** products, including products that are not co-produced by one firm. To get these distances, we use the known distance bounds between co-produced products

[9] One can show that if $\varepsilon_{\omega,\nu}$ are assumed to follow a Frechet distribution with shape parameter α, then this average has an expectation proportional to an average over firms ω of terms of the form $\left((\frac{c_{\omega,\nu}}{\varphi_\omega})^\alpha + (\frac{c_{\omega,\nu'}}{\varphi_\omega})^\alpha\right)^{1/\alpha}$. We remain agnostic as to the empirical distribution of $\frac{c_{\omega,\nu'}}{\varphi_\omega}$ as in a sense that is exactly what the product space reflects in a rich way.

Figure 5: Distances Between Indirectly Connected Products

(a) Max and Min Product Distance (b) Triangle Inequality Distance Bounds



and apply the triangle inequality again.

This idea can be expanded transitively: if Firm A makes products 1 and 3 and Firm B makes products 2 and 3, we can infer bounds on the distances between products 1 and 2 by applying the triangle inequality to the distance bounds between pairs of products 1 and 3 and 2 and 3.[10] We illustrate this idea in Figure 5, mirroring Figure 4: products 1 and 2 are not directly linked; but both of these products are connected to product 3, which allows us to construct bounds for the distance between products 1 and 2.

Note that there may be more than one product, connecting products that are not co-produced. Thus, we find the minimum total length $\overline{d}_{\eta,\eta^0}$ across all of the upper bounds $\overline{d}_{\eta,\nu} + \overline{d}_{\nu,\eta^0}$ for any third product ν. Similarly, we find the greatest length $\underline{d}_{\eta,\eta^0}$ over all lower bounds $\underline{d}_{\eta,\nu} + \underline{d}_{\nu,\eta^0}$. We then define the distance between such indirectly linked varieties as

$$d_{\eta,\eta^0} \equiv \left(\overline{d}_{\eta,\eta^0} + \underline{d}_{\eta,\eta^0}\right)/2.$$

After filling in the distance bounds for indirectly produced products linked by two firms, we iterate this procedure filling in distance bounds for any pair of products that are connected by an arbitrarily long chain of co-production.[11] For groups of products without a chain of co-production (teddy bears and nuclear reactors) then there is no

[10]We take a bounding approach since depending on the underlying assumptions of a model, it is unclear what is the joint distributions of production within a multi-product firm. What may appear as a "stylized fact" of rational firm behavior may simply be generated by random processes often assumed in the literature, see Sheveleva (2019) in the case of multi-product firms and Bernard and Zi (2022) in the case of firm-to-firm networks.

[11]Each iteration is less precise than the previous one, thus making the estimation of sparse distance matrices noisy. We address this challenge in Section 3 by splitting products by broad categories of goods.

basis for assessing distance between these products, so by definition they are in different product clusters.

2.2.2 Assigning Product and Firm Locations

Given a complete set of pairwise distances between products as just constructed, we can represent these products in an $N-1$-dimensional space by assigning each product a vector of length $N-1$. We arbitrarily number products, and set product 1 to zero: $\ell_{\nu_1} = 0$, thus locating product 1 at the origin. For the second product, we choose the first coordinate equal to the distance between products 1 and 2 and the remaining coordinates to zero: $d_{\nu_1,\nu_2} = \|\ell_{\nu_1} - \ell_{\nu_2}\|$. We iterate this procedure for each product i, setting its first $i-1$ coordinates to preserve distances to each product $k < i$: $d_{\nu_i,\nu_k} = \|\ell_{\nu_i} - \ell_{\nu_k}\|$ and the rest to 0.

Second, we locate multi-product firms in the product space.[12] Since firm-product costs $c_{\omega,\nu}/\varphi_\omega$ are observed with errors $\varepsilon_{\omega,\nu}$, the location of firm ω satisfies

$$\|\ell_\nu - \ell_\omega\| = \frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu}.$$

Minimizing the sum of squared errors projects the firm's location onto the hyperplane determined by the varieties it produces $\{\nu_i\}$ and is given by the combination

$$\ell_\omega = \sum_i \frac{c_{\omega,\nu_i}^{\,2}}{\sum_j c_{\omega,\nu_j}^{\,2}} \ell_{\nu_i}. \tag{1}$$

This solution shows that firms are closer to lower cost varieties, after accounting for preferences and competition. Notice also that a firm's proximity to unproduced varieties comes from information embedded in the product space since its location has a weight of zero on unproduced varieties. It follows that in this framework, location embodies each firm's counterfactual costs to produce varieties. Since adopting new products shifts the location of firms, this also implies that product scope contains information on adoption capability.

We now turn to the data and construction of clusters which will form the environment for our estimates of firm behaviour.

[12] Single product firms will have no part in our analysis besides being used in price indexes of competition, both because they may differ in substantial ways from multi-product firms and because they are not informative about co-production patterns.

3 Data and Estimation Procedure

In this section, we present the data and detail how the firm- and product-level distances and coordinates are obtained from the procedure above and then present summary statistics regarding the recovered product space.

3.1 Data

We rely on firm-product-level data from Danish firms, spanning from 2000 to 2018, provided by Denmark Statistics (DST). Specifically, we utilize two data sources: the Production Statistics (VARS) and the Trade Statistics (UHDI). The Production Statistics is a survey in which manufacturing firms with at least 10 employees are required to report their sales in quantities and values for each product they produce. Sales are recorded independently of the market in which the product is sold, thereby including both domestic and export sales. In the Trade Statistics, firms report their exports and imports by product and destination. Products are reported according to the eight-digit level of the Combined Nomenclature (CN) code, with the firm (CVRNR) being the reporting unit. Notice that the CN classification is equivalent to the Harmonized System classification at the 6-digit level.

Our data preparation closely follows Buus et al. (2022), who have provided the code for the estimation of marginal costs using the method proposed by De Loecker et al. (2016). To account for changes in product categories over time, we employ the algorithm proposed by Van Beveren et al. (2012), aggregating categories to the so-called CN8+ level. A product is defined at both the CN8+ and the unit of measurement level. For most CN8+ categories, firms report the same unit of measurement (kg, number, etc.). In some rare instances, the same product code is recorded with different units, and we consider these as separate products.

For estimating marginal costs through production function estimation, we also rely on additional firm registers that provide information on firm-level characteristics such as labor and capital.

3.2 Marginal Cost Estimation

We consider two measures of marginal costs: one based on inverting a Constant Elasticity of Substitution (CES) demand system, and another derived from estimating the production function and markups, as outlined by De Loecker et al. (2016) (DGKP). While the CES method relies on strong assumptions such as constant markups and a particular nesting of demand across varieties, it does not rely on production function estimation or availability of quantities or unit values and is therefore widely applicable. By leveraging sophisticated methods and rich data, the DGKP method allows for economies of scope, scale and selection into multi-product status as well as productivity improvements that accompany production adoption. DGKP therefore represents our preferred estimates for the our product adoption and dropping results. As the CES setting provides a well understood framework to quantify counterfactuals, we will use it for measures of revenue potential, competition and welfare since the DGKP setting does not imply any demand aggregation assumptions.

3.2.1 Consumers and Firms

Recall that there is a discrete number of differentiated products indexed by $\nu = 1, ..., K$ and a finite number of firms indexed by $\omega \in \Omega$. Each firm can produce a variety of each product ν and we index the firm-product variety with subscript ω, ν. We adopt a nested utility function, where preferences over products ν are given by the following Cobb Douglas aggregation:

$$U(\mathbf{q}) = \sum_{\nu} \alpha_{\nu} \ln Q_{\nu} d\nu \quad \text{with} \quad \sum \alpha_{\nu} = 1.$$

Q_{ν} is defined as a CES aggregator over the varieties of product ν supplied in the market as follows:

$$Q_{\nu}(\mathbf{q}) = \left[\sum_{\omega 2 \Omega_{\nu}} q_{\omega,\nu}^{\frac{\sigma \; 1}{\sigma}} \right]^{\frac{\sigma}{\sigma \; 1}} \quad \text{with} \quad \sigma > 1,$$

where $q_{\omega,\nu}$ is the quantity of product ν supplied by firm ω, Ω_{ν} is the set of firms that supply varieties of product ν, and $\sigma > 1$ is the elasticity of substitution.

There is a unit mass of consumers with combined income I. The variety of product ν supplied by firm ω has a price $p_{\omega,\nu}$. Consumers maximize utility through

$$\max_{q_{\omega,\nu}} U(\mathbf{q}) \quad \text{subject to} \quad I = \sum_{\nu} \sum_{\omega 2 \Omega_{\nu}} p_{\omega,\nu} q_{\omega,\nu}.$$

Under a CES utility function, the quantity demanded for a variety of product ν from a firm is given by:

$$q_{\omega,\nu} = \alpha_\nu I / P_\nu^{1-\sigma} p_{\omega,\nu}^{-\sigma}, \tag{2}$$

where P_ν is the price index for product ν that equals:

$$P_\nu \equiv \left(\sum_{\omega \in \Omega_\nu} p_{\omega,\nu}^{1-\sigma} \right)^{1/(1-\sigma)} \tag{3}$$

Notice that total revenues of product ν,R_ν, satisfy the following condition:

$$\alpha_\nu I = P_\nu Q_\nu = R_\nu.$$

3.2.2 Inverting the CES Demand System

By combining the production statistics with the trade statistics, we can compute the total domestic sales (quantity) for each firm-product-year as the difference between the total product value (quantity) and total export value (quantity).[13] Let us define the domestic quantity produced by a firm ω in a product ν (defined as a separate unit-CN8+ code) as $q_{\omega,\nu}$, and the unit value as $p_{\omega,\nu}$.

Assuming monopolistic competition as in Melitz (2003), profit maximization results in constant markups given by $p_{\omega,\nu} = \frac{\sigma}{\sigma-1} MC_{\omega,\nu}$. Consequently, to recover the marginal cost $c_{\omega,\nu}$, we apply the following formula:

$$MC_{\omega,\nu} = \frac{\sigma - 1}{\sigma} P_\nu \left(\frac{p_{\omega,\nu} q_{\omega,\nu}}{R_\nu} \right)^{1/(1-\sigma)} \tag{4}$$

where $R_\nu = \sum_{\omega \in \Omega_\nu} p_{\omega,\nu} q_{\omega,\nu}$ denotes the total sales of product ν by firms.

We calculate the marginal costs for each firm-product by using the domestic unit values to compute the price index (3) and combining it with domestic sales to compute the marginal costs (4). We assign a value of $\sigma = 5$ for all sectors.[14]

[13] We exclude observations with negative domestic sales or quantity, which are likely due to firms engaging in carry along trade (Bernard et al., 2018). This means that firms can export products that they do not produce.

[14] As we only use information on Danish firms and their domestic sales, in computing the marginal cost (4), we implicitly ignore foreign firms exporting to Denmark in the calculation of aggregate revenues and price indexes. This does not pose an issue as we are going to normalize marginal costs $MC_{\omega,\nu}$ by the average marginal cost per product in Section 3.5. In fact, let $P_{\nu DNK}$ denote the price index computed using only the unit values for Danish firms and $R_{\nu DNK}$ denote the aggregate domestic sales of Danish firms. We compute the marginal costs as $MC_{\omega,\nu} = \frac{\sigma-1}{\sigma} P_{\nu DNK} \left(\frac{p_{\omega,\nu} q_{\omega,\nu}}{R_{\nu DNK}} \right)^{1/(1-\sigma)}$. The average marginal cost per product is given by: $\bar{MC}_\nu = \frac{\sigma-1}{\sigma M_\nu} P_{\nu DNK} \left(\frac{1}{R_{\nu DNK}} \right)^{1/(1-\sigma)} \sum_{\omega \in \Omega_\nu} (p_{\omega,\nu} q_{\omega,\nu})^{1/(1-\sigma)}$, where M_ν is the number of firms producing ν. Hence, the normalized marginal cost equals: $\frac{MC_{\omega,\nu}}{\bar{MC}_\nu} =$

3.2.3 Production Function Estimation (DGKP)

We apply the procedure outlined by De Loecker et al. (2016), which integrates the estimation of a production function and markups, to derive marginal costs for each firm-product. In this scenario, there is no need to assume a demand function to determine marginal costs, as these are inferred from the production decisions of firms. We closely follow Buus et al. (2022) in estimating markups and refer readers to their paper for further details.

The key challenge in estimating a production function for multi-product firms arises from a lack of information on how various inputs are allocated to each product. To address this challenge, De Loecker et al. (2016) propose estimating production functions for single-product firms. The production function is estimated using a control function approach that accounts for unobserved heterogeneity in both productivity and input prices. We follow Buus et al. (2022) and also incorporate product-specific export status, number of export destinations, and the square of the number of export destinations into the control functions. However, unlike Buus et al. (2022), we do not include information on export support, as it is only available for a subset of the years.

To calculate markups, this approach uses the results of cost minimization of flexible inputs.[15] Under perfect competition, with no markups, the revenue share of an input and its output elasticity are equal to each other. The markup is defined as the wedge between the revenue share of a variable production input and its output elasticity. By estimating the production function, we determine the output elasticity of materials and we interpret the ratio of this elasticity to the revenue share of materials as the price-cost markup.

Once the markups are obtained, we calculate the marginal cost as the ratio between unit value and markups. To avoid outliers, we calculate the absolute growth rate of firm-product marginal cost changes and exclude the top 2% of those. Moreover, we drop the top and bottom 3% of estimated markups. We amalgamate the CES and DGKP samples to ensure that the estimation of the product space, according to the two different marginal costs, is derived using an identical set of firm-products.

In summary, for each firm-product-year in our sample, we have estimated the marginal costs using either the CES demand function (CES) or a production function estimation (DGKP). The subsequent step in the empirical analysis involves dividing firm-products into clusters of connected products.

$\frac{(p_{\omega,\nu}q_{\omega,\nu})^{1/(1-\sigma)}}{\sum_{\omega\in\Omega_\nu}(p_{\omega,\nu}q_{\omega,\nu})^{1/(1-\sigma)}M_\nu^{-1}}$ and is independent of the price index and aggregate revenues.

[15] For details, see De Loecker and Warzynski (2012) and De Loecker (2021).

3.3 Cluster Construction

To address dimensionality challenges in our analysis and to reduce the lack of precision in estimating a highly sparse product space, we divide our sample into 15 sectors based on CN 2-digit codes.[16] We further refine our selection of products and firms for analysis, as the estimation of the product space is applicable only to a set of products that are either directly or indirectly linked.

When a firm produces both product A and B, the relationship between A and B is defined as a ***direct linkage***. When firm 1 produces both A and C and firm 2 produces both B and C, we define the relationship between product A and C as an ***indirect linkage***. We define a ***cluster*** as a group of products within a sector as the largest set of ***directly or indirectly linked*** products. Identifying clusters is important because we can only estimate distances for products within a cluster and the distance between products in different clusters is infinite by definition. Furthermore, we focus on clusters that are persistent across years and avoid clusters that only occur for a few years due to inconsistent linkages.

We take the following steps in our analysis. First, we exclude single-product firms from our sample, as they do not offer any insight into the links between products or the distance between any two products. For the remainder of the analysis, we exclusively focus on firms that are multi-product. Furthermore, we drop any sector with less than 10 products. This leaves us with 12 sectors.

Following the above definition, a cluster consists of all products with finite distance to each other in ***any*** year within each sector. We further refine the sample of products to avoid the possibility that one cluster in one year breaks down by 2 or more separate clusters, or sub-clusters. If there are more than one sub-cluster per cluster each year, we keep the sub-cluster with the largest number of products in it, and drop the products that are in the other clusters for all years. This procedure drops approximately 15% percent of observations at the firm-product-year level and 19% at the product-year level. For each sector, we find a single cluster.

By definition, each cluster stands distinct from the others. Given this distinction, we will refer to a firm as a firm-cluster for the rest of our analysis. Notably, firms that extend across multiple clusters represent 8% of the total number of firms and account for 23% of overall sales. These are firms whose products span multiple sectors, e.g., a firm making both chemical products and plastic products. In a robustness test, we employ a more

[16]These sectors include: Animal products (CN 2-digit 01-05), Vegetable products (06-15), Foodstuffs (16-24), Mineral products (25-27), Chemical products (28-38), Plastics and rubber (39-40), Leather and Fur (41-43), Wood products (44-49), Textiles (50-63), Footwear and headgear (64-67), Stone and glass (68-71), Metals (72-83), Machinery and electrical (84-85), Transportation (86-89), and Miscellaneous (90-97).

aggregated sector definition, yielding similar outcomes (refer to Appendix D).[17] When dealing with larger clusters, the proportion of firms spanning multiple clusters diminishes, making up 5% of firms and contributing 4% to sales. Furthermore, with larger clusters, the procedure that drops the smallest sub-clusters per cluster each year leads us to drop 8% percent of observations at the firm-product-year level and 12% at the product-year level.

3.4 Cluster Analysis

The descriptive statistics for the clusters we have identified can be found in Table 1. On average, a cluster comprises 88 products and is associated with 53 firms. The distributions of both products and firms exhibit a right skew and display significant interquartile variation. Notably, the counts of products and firms both witnessed a decline around the time of the 2008 financial crisis.

Table 1: Cluster Descriptive Statistics

	Number of Products					Number of Firms				
Year	Avg.	Std.	Med.	25P.	75P.	Avg.	Std.	Med.	25P.	75P.
2000	88	87	53	27	131	62	44	56	26	84
2001	90	83	70	18	138	62	42	54	35	87
2002	91	80	67	23	141	68	40	67	39	95
2003	90	77	62	23	137	61	36	54	39	79
2004	97	77	80	30	150	73	38	68	44	113
2005	94	76	82	18	149	65	41	53	38	107
2006	98	81	89	18	153	60	43	56	20	100
2007	68	58	47	23	102	41	32	32	16	66
2008	73	61	63	22	108	41	27	33	20	63
2009	77	66	71	14	123	44	30	42	18	67
2010	75	63	63	13	113	47	33	36	19	79
2011	83	67	78	23	127	49	36	36	19	78
2012	88	69	84	24	136	46	32	38	21	80
2013	88	71	66	21	141	46	35	32	19	78
2014	91	74	68	16	155	47	36	40	20	78
2015	93	76	66	21	156	46	38	33	17	82
2016	94	73	92	20	151	47	38	37	17	81
2017	97	76	93	18	151	46	39	39	9	79
2018	97	76	98	25	146	48	37	39	17	81
Average	**88**	**73**	**73**	**21**	**137**	**53**	**37**	**44**	**24**	**83**

In each year, we compute average (Avg.), standard deviation (Std.), median (Med.), and 25th and 75th percentiles (25P. and 75P.) of the number of products (first four colums) and number of firms (last four columns) across clusters. In each year, there are 12 clusters (for 12 sectors defined as groups of CN 2-digit codes). The last row (Average) reports the average of the statistics across years.

[17] In this context, the sectors are categorized as: Animals/Vegetables/Food (CN 2-digit 01-24), Minerals/Chemicals/Plastics (25-40), Textiles/Footwear (41-43, 50-67), Stone/Metals (68-83), Machinery/Transportation (84-89), and Miscellaneous (44-49, 90-97).

In Table 2, we present summary statistics that describe the average size of each cluster, both in terms of products and firms. Among these, the Foodstuffs cluster has the highest count of products and firms, while the Transportation cluster registers the lowest. The Metals and Machinery and Electrical clusters both tally above-average numbers in terms of products and firms. Meanwhile, the Textiles cluster stands out with its large product count but a smaller firm count.

Table 2: Cluster Descriptive Statistics

	Number of Products		Number of Firms	
Cluster	Avg.	Std.	Avg.	Std.
Animal Products	106	35	37	19
Foodstuffs	227	19	107	13
Mineral products	21	7	14	8
Chemical Products	148	26	46	11
Plastics and rubber	54	13	41	10
Wood products	19	7	38	25
Textiles	178	33	39	22
Stone and glass	15	5	25	15
Metals	103	23	95	21
Machinery and Electrical	127	16	92	12
Transportation	10	3	10	2
Miscellaneous	48	11	89	34

In each cluster, we compute average (Avg.) and standard deviation (Std.) of the number of products (first two colums) and number of firms (last two columns) across years.

3.5 Distance Estimation

Using the CES demand system inversion and the DGKP approach, we derive two distinct measures for the marginal costs for product ν of firm ω, $MC^i_{\omega,\nu}$, where $i = CES, DGKP$ indexes the methodology. Recall that in our algorithm, we define the marginal cost as $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu}$. We link the marginal costs to $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu}$ using three alternative formulations:[18]

1. Log (Baseline): $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu} = \ln\left(1 + \frac{MC^i_{\omega,\nu}}{\bar{MC}^i_\nu}\right)$. This takes the natural logarithm of one added to the normalized marginal cost.

2. Level: $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu} = \frac{MC^i_{\omega,\nu}}{\bar{MC}^i_\nu}$. Here, we directly use the normalized marginal cost.

3. Inverse Hyperbolic Sine: $\frac{c_{\omega,\nu}}{\varphi_\omega}\varepsilon_{\omega,\nu} = \ln\left(\frac{MC^i_{\omega,\nu}}{\bar{MC}^i_\nu} + \sqrt{1 + \left(\frac{MC^i_{\omega,\nu}}{\bar{MC}^i_\nu}\right)^2}\right)$, another approach maintaining positive distances for marginal costs from the literature.

[18]In our framework we assume that there is a monotone relationship between firm to product distances and corresponding marginal costs. We, however, do not assume that it is necessarily linear and remain agnostic regarding the functional form of this relationship (for example in physics, the force of gravity depends on the inverse of the squared distance).

In each of these formulations, $\bar{MC}_{\nu}^{i}$ idenotes the average marginal cost for product ν across all firms. We normalize our marginal costs by the average to avoid the inevitable unit issues that arise when comparing marginal costs for different products. Namely, it is not obvious whether a higher marginal costs for a kilo of cheese than for a kilo of apples implies that the firm is closer to apples than cheese. By normalizing the marginal costs, a firm is closer to apples if its marginal costs relative to the average competitor are smaller for apples than for cheese. Our primary specification uses the Log formulation.[19] In Appendix C.6, we report the results using marginal costs in levels and in Appendix C.7, those that utilize the inverse hyperbolic sine of marginal costs. our results remain robust across formulations.

We now have all the elements to determine the location of products and firms and the respective distances. For each cluster-year, we construct the product space using the procedure outlined in Section 2, and the three formulations for the marginal costs.

3.6 Distances Based on CN Classification

To evaluate our method against a discrete classification system and to validate the efficacy of this system in predicting product adoption and firm responses to external shocks, we reference an alternate product space derived from the CN classification system. In the CN classification, the distances between firms and products, as well as between firms themselves, are discrete. We assume these distances can have one of four distinct values.

Recall that a product ν is defined as a combination of a unit and CN8+ code. For every firm-cluster in a given year, the core product $\bar{\nu}_{\omega}$ is defined as the product ν with the largest sales for the firm ω. Then, the distance of a firm ω to a product ν can be computed as:

$$d(\omega, \nu) = 1 + \mathbf{1}_{\text{CN6}(\bar{\nu}_{\omega}) \notin \text{CN6}(\nu)} + \mathbf{1}_{\text{CN4}(\bar{\nu}_{\omega}) \notin \text{CN4}(\nu)} + \mathbf{1}_{\text{CN2}(\bar{\nu}_{\omega}) \notin \text{CN2}(\nu)}$$

To elucidate, a firm has a distance of one to all products that fall under the same CN6 code of its core product. It has a distance of two to products under the same CN4 code of its core, but a different CN6 code, and so on.

[19] If we define marginal costs as $\frac{c_{\omega,\nu}}{\varphi_{\omega}} = \ln\left(\frac{MC_{\omega,\nu}^{i}}{\bar{MC}_{\nu}^{i}}\right)$, we would obtain negative values for all the products with a marginal cost below the average. This is problematic because these marginal costs are measures of distances between firms and products and must be positive. Hence, in our baseline specification, we consider $\frac{c_{\omega,\nu}}{\varphi_{\omega}} = \ln\left(1 + \frac{MC_{\omega,\nu}^{i}}{\bar{MC}_{\nu}^{i}}\right)$.

3.7 Distance Analysis

In this section, we present descriptive information about the product space. The summary statistics for the estimated distances in the year 2000 are shown in Table 3. We observe that both product-to-product and firm-to-product distances are smaller in the CES than in the DGKP specification. Additionally, in both specifications, product-to-product distances are consistently larger than firm-to-product distances across the distribution.

Table 3: Estimated Distances: Summary Statistics (2000)

	Product-to-Product		Firm-to-Product	
	CES	DGKP	CES	DGKP
Average	0.39	0.75	0.23	0.58
Std. Dev.	0.23	0.32	0.09	0.21
5th Perc.	0.11	0.27	0.10	0.20
10th Perc.	0.15	0.38	0.13	0.31
25th Perc.	0.23	0.56	0.17	0.46
50th Perc.	0.33	0.69	0.22	0.58
75th Perc.	0.50	0.91	0.28	0.70
90th Perc.	0.70	1.11	0.34	0.81
95th Perc.	0.82	1.29	0.39	0.88

The estimated product space maintain reasonable stability over time. We examine the distribution of the growth rate in the distance between a firm and a product, defined as the difference $d_{\omega\nu t} - d_{\omega\nu t\ \ 1}$, since the distances are expressed in logarithms. Both the average and the median firm-to-product growth rate of distance is approximately zero a year to year basis. In the CES framework, distances vary by 5 percentage points within the 25-75 percentile range, indicating general stability of our approach. Distances in the DGKP are slightly more volatile, as distances vary by 14 percentage points with the 25-75 percentile range. We provide these results in Table C.1 of the appendix.

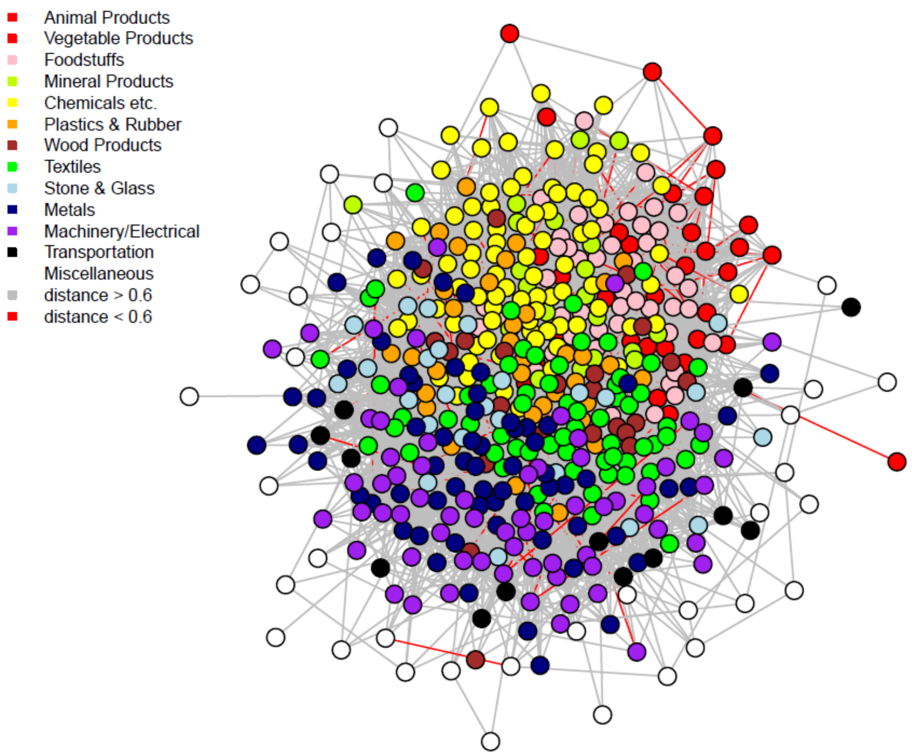
Due to the high dimensionality of the product space and to confidentiality constraints, we are unable to provide a visual representation of the product and firm space at the CN8 level. We can however provide such a representation with aggregation to the CN4 level using the DGKP approach with log cost coordinates.[20] Our clustering algorithm produces a single cluster for the entire set of CN 4-digit products. This means that each CN 4-digit product is directly or indirectly linked to all others. Figure 6 provides a network representation of the product space using maximum-spanning trees, constructed following methods similar to those in Hidalgo et al. (2007).[21] The figure illustrates that products

[20]We compute marginal costs as a quantity-weighted average for each CN 4-digit product-unit as the quantity-weighted average of the CN 8-digit product-units within that CN 4-digit product-unit.

[21]Lines are firm drawn to create the smallest tree including all products. Subsequently, lines below a certain cutoff distance are added to the graph. Some distances are omitted to enhance readability.

within similar sectors tend to cluster together, such as Chemical products alongside Plastic and Rubber, or Machinery in close proximity to Metals and Transportation. However, it is important to observe the diversity within these clusters: a typical metal product is equidistant from other metal products as it is from machinery or transportation items. Additionally, the periphery of the space is marked by an assortment of miscellaneous products.

Figure 6: Product Space



However, to offer some insight into the relationship between distances within this space and observable characteristics of products and firms, we conduct several regression analyses, using the sample of products defined as CN 8-digit product-units. We find that products with higher sales are generally more isolated than those with smaller sales. Moreover, we observe a hump-shaped relationship between product-to-product distances and co-production, suggesting that products are further apart when only a few firms co-produce them. However, as the number of co-producing firms increases beyond certain thresholds, the distance between products starts to decrease. Moreover, we find that firms tend to be closer to the products they produce, which acts as a validation of our approach. Finally, firms with larger sales tend to be closer to their products. Details are in Appendix C.1.

We now move to our estimates of firm behaviour with counterfactual measures from the product space.

4 Product Adoption and Proximity

While firm-level shocks, such as a boost in productivity, and product-level shocks, like rising demand, might prompt a firm to introduce a new product or discontinue an existing one, the precise prediction of which product will be introduced or dropped has largely been overlooked in the literature.[22] By estimating a product space, we can enhance standard models to identify which products are more likely to be introduced, based on their positional relationship to the firm. When a firm opts to introduce a new product from among those it does not currently produce, it is likely to select a product that is proximal in the space. This section first quantifies the predictive capability of the product space in determining which products will be introduced. In Appendix C.2, we also examine the predictive capability regarding product discontinuations. The section continues with an IV strategy to test whether demand shocks at the product level attract closer by firms to adopt them. The results show that this is indeed the case.

4.1 Product Adoption by Distance Rank

For each firm, we select products not produced in its initial year within the dataset. For example, if a firm is first included in the dataset in 2000, our sample consists of products in the firm's cluster that it does not produce that year. In subsequent years after entry, we compute a production indicator $\text{Intro}_{\nu\omega ct}$ for product ν, firm ω, cluster c, and year t, equal to 1 if the product is produced by the firm and 0 otherwise. This value is then normalized by the average product introduction rate, approximately 0.4 percent.

We first estimate the following equation:

$$\text{Intro}_{\nu\omega ct} = \beta Rank_{\nu\omega ct\ \ 1} + a_{\omega t} + b_{\nu t} + \epsilon_{\nu\omega ct} \tag{5}$$

where $Rank_{\omega,\nu ct\ \ 1}$ represents the rank of products not produced by firm ω, based on distance in the previous year $d_{\omega,\nu ct}$, such that the closest product k satisfies $Rank_{kfct} = 1$.[23] We include firm-year fixed effects ($a_{\omega t}$) to control for any shock at the firm-level that

[22]For instance, standard models that base a firm's product mix decision on a core competence predict that a positive productivity shock will expand a firm's scope, and that the new product will be far removed from the core. However, such models cannot predict ex-ante whether the new product introduced will be product A or product B. Ex-post, they assume that if product A is introduced, it must be closer to the firm's core than product B.

[23]We calculate a unique ranking of products. In the event that two products have the same distance from the firm, they are randomly ranked by the code. Notably, this is only applicable to the four distances based on the CN classification: all products within the same aggregation have identical distances to a

would cause product adoption and product-year fixed effects ($b_{\nu t}$) to control for shocks to demand or technology that might affect a product in a year. By definition, each firm and product is exclusive to a single cluster, implying that the set of fixed effects inherently controls for cluster-year shocks. The findings are illustrated in Table 4.

Table 4: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.005***	-0.005***	-0.005***
	(0.001)	(0.001)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.07
# Obs.	645290	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

The columns demonstrate that product adoption rates diminish as the distance ranking from a firm increases across all three specification measures. In terms of magnitude, transitioning from the M-th closest product to the M+1-th closest reduces the relative probability of product introduction by 0.5 percentage points for CES and DGKP.[24]

Discrete VS Continuous Classifications. As illustrated in Table 4, using definitions of distances based on the CN classification yields results that are comparably effective in predicting product introduction to those produced using the continuous classification we propose. The findings imply that the CN classification offers valuable insights by aggregating products into categories, given that firms proximate to these categories exhibit a higher likelihood of introducing these products.

Table 5: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Rank (=1 closest)	-0.002***	-0.003***
	(0.001)	(0.000)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

firm.

[24]For the CN classification, discreteness means we assign random ranks within a given distance, and the result above shows that the classification does contain meaningful information about product proximity.

Nevertheless, the CN classification, while valuable, presents a notable disadvantage compared to a continuous measure due to its discrete nature, lacking the ability to differentiate between products within the same category. For instance, per the CN classification, a firm whose core product is in cotton fabrics would be deemed equally likely to introduce apparel and clothing accessories as it would be to introduce carpets and other textile floor coverings, since apparel and carpets belong to different CN 2-digit codes relative to cotton, but within the same cluster. Reiterating the regression (5), incorporating CN distance fixed effects as in Table 5 affirms that the continuous classification approach predicts product adoption even within CN hierarchies, as the coefficients on $Rank_{\omega,\nu ct-1}$ are negative and statistically significant for both CES and DGKP.[25]

4.2 Product Adoption by Distance

We next estimate the following equation:

$$\text{Intro}_{\nu\omega ct} = \beta d_{\nu\omega ct-1} + a_{\omega t} + b_{\nu t} + \epsilon_{\nu\omega ct} \tag{6}$$

where $d_{\nu\omega ct-1}$ represents the distance of product ν to firm ω in the previous year and, in our baseline specification, is already estimated as a natural logarithm. Similarly to the previous section, We include firm-year fixed effects and product-year fixed effects. The findings are illustrated in Table 6.

Table 6: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Distance	-3.685***	-1.543***	-1.688***
	(0.481)	(0.171)	(0.042)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.08
# Obs.	645290	645290	645290

Results from OLS estimation of (6). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

The columns demonstrate that product adoption rates diminish as the distance ranking from a firm increases across all three specification measures. In terms of magnitude, doubling the log distance reduces the probability of product adoption by 3.6 percentage points in the CES specification, by 1.5 percentage points in the DGKP specification, and by 1.7 percentage points in the CN specification, or an increase of 2.6 pp for CES and

[25]Since the rank within CN distance bins is determined randomly, we omit displaying results for distances based on the CN classification.

Table 7: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Distance	-1.615***	-0.737***
	(0.481)	(0.171)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (6). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

1.1 pp for DGKP when costs are halved. Reiterating the regression (5), incorporating CN distance fixed effects as in Table 7 again affirms that the continuous classification approach predicts product adoption even within CN hierarchies.

We now turn to a validation exercise by combining product distance with instrumented demand shocks.

4.3 Product Adoption and Proximate Demand Shocks

As highlighted in the preceding section, a myriad of reasons can prompt a firm to introduce a new product–these can span from firm-specific factors, like enhancements in efficiency, to product-specific elements, such as changes in demand. Here we examine how a positive demand shock influences the likelihood of a product being introduced, and how this relationship is modulated by the product's proximity to the firm in the presence of a demand shock.

To quantify a demand shock that is credibly exogenous to supply conditions, we instrument for positive export demand shocks on product introduction. We estimate

$$\begin{aligned}\text{Intro}_{\omega,\nu ct} =& \beta_1 \text{Log Distance}_{\omega,\nu ct\ \ 1} + \beta_2 \text{Log Exports}_{\nu t} \times \text{Log Distance}_{\omega,\nu ct\ \ 1} \\ &+ FE_{\nu t} + FE_{\omega t} + FE_{\omega,\nu} + \epsilon_{\omega,\nu ct}\end{aligned} \tag{7}$$

where the dependent variable $\text{Intro}_{\omega,\nu ct}$ mirrors that used in regression (5); it is assigned a value of one if product ν is produced by firm ω in cluster-year ct. Notice that our sample is constrained to products that are not produced in the initial year that a firm appears in our data. For example, if a firm makes its initial appearance in the dataset in the year 2000, our considered product sample comprises those within the firm's cluster that are not produced in that year. $\text{Log Distance}_{\omega,\nu ct\ \ 1}$ is the lagged log[26] distance between product

[26]Notice that in our baseline specification, we use the log of marginal costs and hence, the distances

ν and firm ω and Log Exports$_{\nu t}$ is the log of the total exports from Denmark of product ν in year t. We include product-year and firm-year fixed effects as in (5). Moreover, this specification encompasses firm-product fixed effects and we restrict the sample to firm-products that have been introduced in the years considered, thereby excluding firm-products ω, ν such that $\max_t \text{Intro}_{\omega,\nu ct} = 0$. Therefore, the identifying variation is in the switch of a product from being not produced to being produced.

We instrument Log Exports$_{\nu t}$ by adhering to the approach by Hummels et al. (2014), similar to Dhyne et al. (2021), using either global exports or those from comparable nations to instrument for Danish exports by calculating Log Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$, with the total exports of all countries except Denmark.[27] Notice that export data, sourced from BACI, is reported at the 6-digit level, while our products are defined at a more granular 8-digit level.

Table 8: Export Shocks and Product Adoption

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	47.600	401.482***	25.540	96.813***	3.720	8.426
	(70.531)	(128.156)	(18.852)	(31.830)	(5.525)	(9.595)
(Lagged Log Distance)X(Log Exports)	-10.772*	-41.703***	-3.429**	-9.555***	-0.561	-0.971
	(6.050)	(11.139)	(1.596)	(2.721)	(0.472)	(0.830)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782
F-Stat		2890.39		3611.69		3284.21

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Exports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$ is the total exports of all countries except Denmark.

In Table 8, we show that our key parameter of interest, the interaction between distance and exports, is negative and statistically significant under the CES and DGKP cost specifications. However, the CN-based distance measures do not attain statistical significance. This outcome suggests that the impact of an export demand shock - to increase the probability of product introduction - is amplified for products proximate to the firm (where potential marginal costs are low) and attenuated for more distant products. When demand increases for a product, closer firms in the product space supply that demand.

obtained with our algorithm are already in logs.

[27] As a robustness, we consider Log Exports IV$_{\nu t}$ = Log $\sum_{k \in K}$ Exports$_{k\nu t}$ where K denotes two different sets of countries: the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA, and a set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden. Results are in appendix C.3. In Appendix C.3, we report the results of the first stage regressions.

Import Shocks and Product Adoption. In a supplementary analysis, we explore the implications of import shocks on product adoption, employing a methodology analogous to that above. Here, the key variable is the interaction between the log of imports and the distance from the firm to the product. Mirroring findings from the export scenario, we observe that the impact of import shocks diminishes across the product space: the more distant a product is from a firm, the less pronounced the effect of import shocks becomes. Comprehensive results are provided in Appendix C.5.

5 Product Distances, Firm Behaviour and Welfare

In this section, we fully utilize the novel hypothetical marginal costs of potential products derived above to investigate the impact of firm proximity to unproduced products. We introduce two new key measures. ***Revenue Potential***, which quantifies latent revenues for each firm in unproduced products akin to Market Potential in the trade literature (Redding and Venables, 2004); and ***Competition Potential***, which quantifies potential losses from competitors producing the current products of each firm. Since these measures hinge upon hypothetical revenues that firms might generate when introducing a new product, we necessitate an assumption regarding the demand structure and market competition, given that our algorithm exclusively returns marginal costs. We adopt a CES demand system as above, wherein the elasticity of substitution is equal to $\sigma = 5$ and firms are monopolistically competitive.[28] Since the distances based on the CN classification are arbitrary and do not correspond to marginal costs, in this section, we apply the described approach only to the CES and DGKP distance metrics, to estimate the impact on sales growth, scope growth and product mix. The section ends with a counterfactual analysis of consumer welfare if all firms in a cluster competed in all varieties.

5.1 Revenue Potential and Competition Potential

In our log-specification, each product ν and firm ω is assigned a marginal cost $c_{\omega,\nu} = \exp(d_{\omega,\nu}) - 1$, applicable to all products within a cluster-time–both those produced and those unproduced by the firm. We omit the time subscript for notational simplicity. Assuming monopolistic competition (implying constant markups), we calculate the revenues

[28] This is the standard setting for much of monopolistic competition literature which implies constant markups, and the implications for aggregate behaviour and welfare when markups are variable, as in the DGKP case, are of course different (Zhelobodko et al., 2012; Dhingra and Morrow, 2019). For our purposes here, one likely difference between the CES and DGKP approaches is that larger firms likely have higher market power and markups (Hottman et al., 2016), which will imply that products produced by larger firms in the DGKP case will appear to have lower costs than in the CES case. Further work could use our measures to inform models of strategic competition to accommodate this.

for existing products as follows:

$$r_{\omega,\nu} = \left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1} R_\nu \left(\frac{c_{\omega,\nu}}{P_\nu}\right)^{1-\sigma}, \tag{8}$$

where R_ν represents the total expenditures on product ν, sourced from data as the total domestic revenues for a product, and P_ν is the price index, computed as:

$$P_\nu = \frac{\sigma}{\sigma-1}\left(\sum_{\omega\in\Omega_\nu} c_{\omega,\nu}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}, \tag{9}$$

with Ω_ν denoting the set of firms producing product ν.

To compute the revenues from potential products for a firm ω, we apply the formula delineated in (8), making adjustments to the price index to include the hypothetical marginal cost for product ν from firm ω. Specifically, let $\tilde{r}_{\omega,\nu}$ represent the hypothetical revenues of firm ω in product ν. Then, these are given by:

$$\tilde{r}_{\omega,\nu} = \left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1} R_\nu \left(\frac{c_{\omega,\nu}}{\tilde{P}_{\omega,\nu}}\right)^{1-\sigma}, \tag{10}$$

where $\tilde{P}_{\omega,\nu}$ denotes the price index adjusted by production from firm ω as:

$$\tilde{P}_{\omega,\nu} = \frac{\sigma}{\sigma-1}\left(c_{\omega,\nu}^{1-\sigma} + \sum_{\omega'\in\Omega_\nu} c_{\nu\omega'}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}. \tag{11}$$

Let V_ω denote the set of products that firm ω produces and W_ω the set of products that firm ω does not produce. We define Revenue Potential (RP) as follows:

$$RP_\omega \equiv \frac{\sum_{\nu\in W_\omega} \tilde{r}_{\omega,\nu}}{\sum_{\nu\in V_\omega} r_{\omega,\nu} + \sum_{\nu\in W_\omega} \tilde{r}_{\omega,\nu}} \tag{12}$$

Some firms, while proximate to large and attractive markets, do not (yet) produce the corresponding goods. This implies that such firms possess substantial potential to augment their revenue, particularly in the event of a positive productivity shock or reduced fixed costs, relative to a firm distanced from all products it does not produce. Consequently, a higher RP is associated with elevated potential revenues as the firm broadens its scope.

Next, we turn our attention to a measure accounting for the effects of potential competition on an individual firm. Assume that all firms within a cluster manufacture a product

ν that is within the scope of firm ω. The revenues of firm ω can then be expressed as:

$$\hat{r}_{\omega,\nu} = \left(\frac{\sigma - 1}{\sigma}\right)^{\sigma - 1} R_{\nu} \left(\frac{c_{\omega,\nu}}{\tilde{P}_{\nu}}\right)^{1 - \sigma}, \tag{13}$$

where $\tilde{P}_{\nu}$ is the price index, calculated under the assumption that all firms in the cluster produce product ν:

$$\tilde{P}_{\nu} = \frac{\sigma}{\sigma - 1} \left(\sum_{\omega \in \Omega} c_{\omega,\nu}^{1 - \sigma}\right)^{\frac{1}{1 - \sigma}}, \tag{14}$$

with Ω representing the set of firms in the cluster.

We introduce a Competition Potential (CP) index, defined as one minus the ratio of a firm's revenue under a hypothetical worst-case scenario in which all firms in the cluster opt to compete with firm ω to the actual revenue of firm ω:

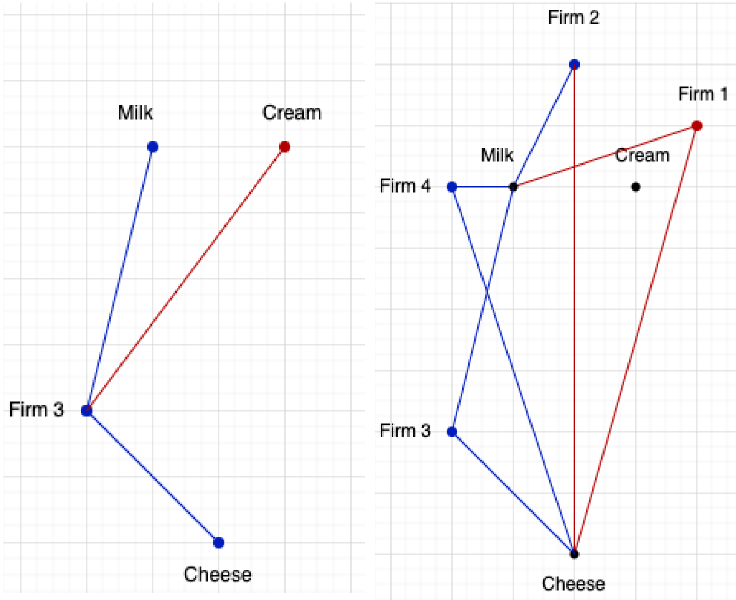
$$CP_{\omega} = 1 - \frac{\sum_{\nu \in V_{\omega}} \hat{r}_{\omega,\nu}}{\sum_{\nu \in V_{\omega}} r_{\omega,\nu}} \tag{15}$$

Hence higher Competition Potential corresponds to hypothetically diminished revenues.

Two stylized representations of Revenue Potential and Competition Potential are depicted in Figure 7a and Figure 7b for Firm 3. Observed production by firms is in blue, while potential production is in red. In Figure 7a, Firm 3 can potentially earn revenues in Cream production, while in Figure 7b, Firm 1 can potentially compete with Firm 3 in Milk and Cheese, while Firm 2 might also additionally compete in Cheese.

Figure 7: Graphical Representation of RP and CP

(a) Revenue Potential (b) Competition Potential



5.2 Sales Growth, Scope Growth and Core Focus

Next, we aim to evaluate whether RP and CP can predict firm performance. We examine the following regression:

$$y_{\omega ct} = \beta_1 RP_{\omega ct\ \ 1} + \beta_2 CP_{\omega ct\ \ 1} + FE_{ct} + FE_{\omega} + \epsilon_{\omega ct} \tag{16}$$

where $y_{\omega ct}$ represents three performance variables for firm ω in cluster c at year t, comprising: 1) the growth rate of domestic sales, $\ln \text{Sales}_{\omega ct} - \ln \text{Sales}_{\omega ct\ \ 1}$, with domestic sales sourced from the data, 2) the growth rate of the scope, $\ln \text{Scope}_{\omega ct} - \ln \text{Scope}_{\omega ct\ \ 1}$, and 3) the Theil T-index of sales concentration within a firm.[29] The Theil Index quantifies the sales concentration within a firm: a higher index indicates higher concentration, implying an elevated focus on the firm's core product (Mayer et al., 2014). $RP_{\omega ct\ \ 1}$ and $CP_{\omega ct\ \ 1}$ are the lagged values of equations (12) and (15), respectively and we include cluster-time and firm fixed effects.

[29] $Theil_{\omega ct} = \frac{1}{\text{Scope}_{\omega ct}} \sum_{\nu 2 V_{\omega}} \left(\frac{\text{Sales}_{\nu \omega ct}}{\text{Average Sales}_{\omega ct}} \right) \ln \left(\frac{\text{Sales}_{\nu \omega ct}}{\text{Average Sales}_{\omega ct}} \right)$.

Table 9: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	(CES)	(DGKP)
Lagged RP	0.525***	0.317***
	(0.072)	(0.055)
Lagged CP	-0.077*	-0.024
	(0.044)	(0.030)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.16
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table 10: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	(CES)	(DGKP)
Lagged RP	0.005	0.058***
	(0.020)	(0.017)
Lagged CP	-0.124***	-0.060***
	(0.016)	(0.012)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Tables 9 and 10 show that a higher Revenue Potential is associated with higher growth rates of sales and of scope. In fact, the coefficient on RP is positive and statistically significant in all specifications. As we control for firm fixed effects, the interpretation is that the growth rate of sales and scope is higher, relative to the average firm growth rate, in the presence of initial higher revenue potential. In contrast, the negative and significant coefficients for CP in Table 10 indicate that higher potential competition is associated with lower scope growth.

Table 11: Firm Core Focus

	Dependent Variable: Theil Index (Core Focus)	
	(CES)	(DGKP)
Lagged RP	-0.023	-0.061***
	(0.025)	(0.019)
Lagged CP	0.031*	0.031**
	(0.016)	(0.012)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.77	0.77
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table 11 shows that higher revenue potential is associated with less focus on the core, while higher competitive potential is associated with higher focus on the core.[30] This is consistent with the results of Mayer et al. (2014), as firms focus on their core products when facing tougher potential competition and focus more on the periphery products when facing higher revenue potential.

5.3 Counterfactual: Gains from New Varieties

Here we use our hypothetical marginal costs to answer the following question: how do price levels change when all firms manufacture all products? This scenario implies that all firms, even those that currently do not produce a particular product, experience a positive shock–such as a reduction in fixed costs of production per product–that induces them to introduce such a product at the hypothetical marginal costs we have estimated.

To address this question, we define the Entry Potential (EP) for product ν as the ratio of the price index when all firms in the cluster choose to produce product ν ($\tilde{P}_\nu$, defined in (14)) to the actual price index (P_ν, defined in (9)):

$$EP_\nu = \frac{\tilde{P}_\nu}{P_\nu}. \tag{17}$$

The difference between the two price indexes is driven solely by the new varieties and the counterfactual pattern of costs. Thus, EP provides a measure of the gains from new varieties (Feenstra, 1994). EP is computed for every product ν and year t. Subsequently, we compute a weighted average and weighted standard deviation of EP_ν for each cluster

[30] Note that our model with CES preferences and monopolistic competition can feature changes in the distribution of sales within firms because each product ν is its own nest, with its own price index. Under CES preferences and monopolistic competition, the Theil Index would depend on the distribution of marginal costs within the firm if the products are differentiated varieties of the same good.

c in each year, where the weights are the expenditure shares on each product ν. In Table 12, we report the average and standard deviation of EP_ν across years for each cluster.[31]

Table 12: Entry Potential by Sector ($\times 100$)

	CES		DGKP	
Cluster	Avg.	Std.	Avg.	Std.
Animal Products	72.3	21.7	75.0	23.7
Foodstuffs	69.6	19.0	77.6	25.4
Mineral products	73.0	24.8	77.9	20.9
Chemical Products	77.0	21.4	81.3	26.8
Plastics and rubber	75.4	22.8	80.9	27.1
Wood products	79.3	18.7	91.0	16.0
Textiles	70.2	15.5	74.0	23.1
Stone and glass	80.5	26.2	87.3	21.8
Metals	78.6	23.9	84.0	27.2
Machinery and Electrical	70.8	22.5	77.9	29.3
Transportation	82.2	23.0	90.9	19.2
Miscellaneous	83.2	15.4	91.8	16.6
Average	**76.0**	**21.2**	**82.5**	**23.1**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index.

On average, when all firms manufacture a variety of each product, the price index falls to 76% of its initial value for the CES marginal costs, with Foodstuffs recording the largest drop (69.6%) and Miscellaneous the smallest (83.2%). Using DGKP marginal costs results in a smaller drop in the price index (to 82.5%).

These results suggest that if existing firms start producing new products, it can improve consumer welfare. At the same time, these welfare gains are limited, suggesting that on average, most firms with high capabilities to produce a given product already produce it.

5.4 Entry Potential for Green Products

As a sample application, we compute the Entry Potential for a list of 6-digit "green products" provided by Mealy and Teytelboym (2023), based on the work by Mealy and Teytelboym (2022) and Andres et al. (2023). Out of the 295 products provided in the original list, only 123 are included in our clusters. The remaining products that are not matched are either not produced in Denmark or excluded from our clusters given the algorithm we outlined above. The results in Table 13 are in line with our baseline results: if all firms begin to make all of the green products, the price index for these products

[31] Since the gains from new varieties depend on the elasticity of the price index with respect to the number of varieties, which is a function of σ, utilizing a smaller value of σ yields larger gains.

would be 80-85% that of the current one. This indicates that consumer surplus gains are achievable by subsidizing product introduction in this sector.

Table 13: Entry Potential by Sector

	CES		DGKP	
	Avg.	Std.	Avg.	Std.
Green Products	0.80	0.22	0.85	0.26

Average and standard deviation of the EP for green products. We set $\sigma = 5$ in the calculation of the price index.

6 Conclusion and Future Directions

Using observed production patterns within and across firms, we construct a continuous, high-dimensional product classification or ***product space***. By locating firms in this space, we reveal how proximity to potential products or competitors shapes product adoption, sales and scope growth, and core focus. Distance to potential products explains which products a firm will adopt, tempered by local competition characteristics. Letting the data reveal such a classification system allows us to discover new dimensions of firm behavior with respect to both products and other firms. The distance rank of potential products away from firms helps explain the path of product adoption and a Hummels et al. (2014) style export demand instrument shows that new demand is supplied by closer firms. Revenue Potential and Competition Potential have consistent and significant impacts on sales and scope growth in addition to core focus. By measuring Entry Potential, we also find that consumer surplus could increase by 10-30% if all firms competed in all products connected by co-production by using counterfactual costs implied by distance.

While our novel methodology affords a new perspective on multi-product firm activity, applications of our approach span beyond the questions considered here. Some exciting potential applications of our method are to explain behaviour across global markets as this micro foundation naturally extends to using transport and distance measures from the international trade literature. While we have restricted ourselves to product dynamics, a spatial analysis of the product space might have implications for entry and exit dynamics of firms, extending ideas from which products are chosen to which firms are selected. In addition, a complementary demand side product classification could be produced based on household consumption data, rather than production and sales patterns. While only a first pass at understanding growth strategies and competition in a micro founded product space constructed from industry wide data, it does so without relying on handed down classifications of economic activities. It has the potential to free subsequent analysis from some of the vagaries of categorization systems across countries and over time.

References

Andres, P., P. Mealy, N. Handler, and S. Fankhauser (2023): "Stranded nations? Transition risks and opportunities towards a clean economy," *Environmental Research Letters*, 18, 045004.

Arkolakis, C., S. Ganapati, and M.-A. Muendler (2021): "The Extensive Margin of Exporting Products: A Firm-level Analysis," *American Economic Journal: Macroeconomics*, 13.

Autor, D., D. Dorn, and G. H. Hanson (2013): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *The American Economic Review*, 103, 2121–2168.

Bernard, A. B., E. J. Blanchard, I. Van Beveren, and H. Vandenbussche (2018): "Carry-Along Trade," *The Review of Economic Studies*, 86, 526–563.

Bernard, A. B., S. J. Redding, and P. K. Schott (2010): "Multiple-Product Firms and Product Switching," *American Economic Review*, 100, 70–97.

——— (2011): "Multiproduct Firms and Trade Liberalization," *The Quarterly Journal of Economics*, 126, 1271–1318.

Bernard, A. B. and Y. Zi (2022): "Sparse production networks," Tech. rep., National Bureau of Economic Research.

Bishop, A., J. Mateos-Garcia, and G. Richardson (2022): "Using text data to improve industrial statistics in the UK," *Working Paper*.

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81, 1347–1393.

Boehm, J., S. Dhingra, and J. Morrow (2022): "The comparative advantage of firms," *Journal of Political Economy*, 130, 3025–3100.

Buus, M. T., J. R. Munch, J. Rodrigue, and G. Schaur (2022): "Do export support programs affect prices, quality, markups and marginal costs? Evidence from a natural policy experiment," *Review of Economics and Statistics*, 1–45.

De Loecker, J. (2021): "Comment on *(Un)pleasant...* by Bond et al (2020)," *Journal of Monetary Economics*.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016): "Prices, Markups, and Trade Reform," *Econometrica*, 84, 445–510.

De Loecker, J. and F. Warzynski (2012): "Markups and Firm-Level Export Status," *The American Economic Review*, 102, 2437–2471.

Dhingra, S. (2013): "Trading Away Wide Brands for Cheap Brands," *American Economic Review*, 103, 2554–84.

Dhingra, S. and J. Morrow (2019): "Monopolistic competition and optimum product diversity under firm heterogeneity," *Journal of Political Economy*, 127, 196–232.

Dhyne, E., A. K. Kikkawa, M. Mogstad, and F. Tintelnot (2021): "Trade and Domestic Production Networks," *The Review of Economic Studies*, 88, 643–668.

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

Eckel, C. and P. J. Neary (2010): "Multi-Product Firms and Flexible Manufacturing in the Global Economy," *Review of Economic Studies*, 77, 188–217.

Escolar, E. G., Y. Hiraoka, M. Igami, and Y. Ozcan (2023): "Mapping firmsâ locations in technological space: A topological analysis of patent statistics," *Research Policy*, 52, 104821.

Feenstra, R. and H. Ma (2007): "Optimal Choice of Product Scope for Multiproduct Firms under Monopolistic Competition," *in E. Helpman, D. Marin and T. Verdier, eds., The Organization of Firms in a Global Economy, Harvard University Press*.

Feenstra, R. C. (1994): "New Product Varieties and the Measurement of International Prices (Appendix)," *American Economic Review*, 84, 157–177.

Flach, L. and M. Irlacher (2018): "Product versus Process: Innovation Strategies of Multiproduct Firms," *American Economic Journal: Microeconomics*, 10, 236–77.

Girard, M. and A. Trau (2004): "Implementing the North American Industry Classification System: The Canadian Experience," Tech. rep., Statistics Canada.

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010): "Multiproduct firms and product turnover in the developing world: Evidence from india," *The Review of Economics and Statistics*, 92, 1042–1049.

Grant, M. (2023): "Where Do the Data Come From? Endogenous Classification in Administrative Data," *Working Paper*.

Hausmann, R., J. Hwang, and D. Rodrik (2007): "What you export matters," *Journal of Economic Growth*, 12, 1–25.

Hidalgo, C. A., B. Klinger, A.-L. Barabàsi, and R. Hausmann (2007): "The Product Space Conditions the Development of Nations," *Science*, 317, 482–487.

Hoberg, G. and G. Phillips (2016): "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 124, 1423–1465, publisher: University of Chicago Press Chicago, IL.

Hossain, A. M. and W.-S. Soh (2015): "A survey of calibration-free indoor positioning systems," *Computer Communications*, 66, 1–13.

Hottman, C. J., S. J. Redding, and D. E. Weinstein (2016): "Quantifying the sources of firm heterogeneity," *The Quarterly Journal of Economics*, 131, 1291–1364, publisher: Oxford University Press.

Hummels, D., R. Jorgensen, J. Munch, and C. Xiang (2014): "The wage effects of offshoring: Evidence from Danish matched worker-firm data," *American Economic Review*, 104, 1597–1629, publisher: American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Irlacher, M. (2022): "Multi-Product Firms in International Economics," *CESifo Working Paper*.

Jacobs, G. and C. ONeill (2003): "On the reliability (or otherwise) of SIC codes," *European Business Review*, publisher: MCB UP Ltd.

Jaffe, A. (1986): "Technological Opportunity and Spillovers of R&D: Evidence from Firms, Patents, Profits and Market Value," *American Economic Review*, 76, 984–1001.

Jakel, I. C., C. Eckel, L. Macedoni, and R. Riezman (2023): "Explorations in Product Space: Multi-Product Firms and their Product Portfolios," *Working Paper*.

Kogan, L., D. Papanikolaou, L. D. Schmidt, and B. Seegmiller (2021): "Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations," Tech. Rep. 29552, National Bureau of Economic Research.

Macedoni, L. (2022): "Large multiproduct exporters across rich and poor countries: Theory and evidence," *Journal of Development Economics*, 156, 102835.

Macedoni, L. and M. J. Xu (2022): "Flexibility and Productivity: Towards the Understanding of Firm Heterogeneity," *International Economic Review*.

Mayer, T., M. J. Melitz, and G. I. Ottaviano (2014): "Market Size, Competition, and the Product Mix of Exporters," *The American Economic Review*, 104, 495–536.

Mealy, P. and A. Teytelboym (2022): "Economic complexity and the green economy," *Research Policy*, 51, 103948.

——— (2023): "Green Transition Navigator," .

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725.

Nocke, V. and S. Yeaple (2014): "Globalization and multiproduct firms," *International Economic Review*, 55, 993–1018.

Pellegrino, B. (2019): "Product differentiation and oligopoly: A network approach," *WRDS Research Paper*.

Redding, S. and A. J. Venables (2004): "Economic geography and international inequality," *Journal of International Economics*, 62, 53–82.

Sheveleva, L. (2019): "Multi-product Exporters: Facts and Fiction," *SSRN Electronic Journal*.

Van Beveren, I., A. Bernard, and H. Vandenbussche (2012): "Concording EU Trade and Production Data over Time," NBER Working Papers 18604, National Bureau of Economic Research, Inc.

Wu, H., C. Wang, and N.-F. Tzeng (2005): "Novel self-configurable positioning technique for multihop wireless networks," *IEEE/ACM transactions on networking*, 13, 609–621.

Zhelobodko, E., S. Kokovin, M. Parenti, and J.-F. Thisse (2012): "Monopolistic Competition: Beyond the Constant Elasticity of Substitution," *Econometrica*, 80, 2765–2784.

A Product Classifications Across HS, SIC and NAICS

Table A.1 provides example products from the Harmonized System which vary by main classification across two North American classification systems, the SIC and NAICS and even change main classification over time in some instances.

Table A.1: Product Classification Differences

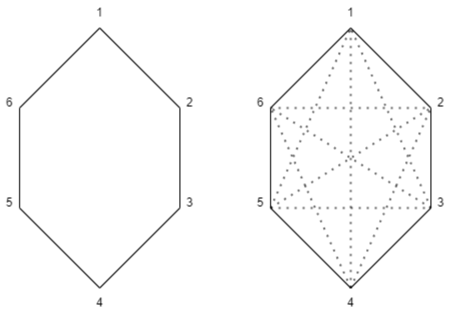
	HS		SIC		NAICS	
	Code	Description	Code	Description	Code	Description
1	1005904040	Popcorn, Unpopped, Except Seed	119	Cash Grains, Not Elsewhere Classified-Con. (Major Group 01 Agricultural production-crops DIVISION-A Agriculture, Forestry, and Fishing)	Before 2014: 111150 After 2014: 311999	111150 Corn Farming (111 Crop Production Sector) 311999 All Other Miscellaneous Food Manufacturing (311-Food Manufacturing)
2	714101000	Cassava (manioc) frozen	2037	Frozen Fruits, Fruit Juices, and Vegetables (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	111130	Dry Pea and Bean Farming (111 Crop Production Sector)
3	1504102000	Fish-liver oils and their fractions	2077	Animal and Marine Fats and Oils (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	114111	Finfish Fishing (114 Fishing, Hunting and Trapping Sector)
4	2301200010	Flours, meals and pellets, of meat or meat offal; greaves	2077	Animal and Marine Fats and Oils (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	114111	Finfish Fishing (114 Fishing, Hunting and Trapping Sector)
5	1702202210 1702202290 1702202410 1702202490 1702202810 1702202890 1702204010 1702204090	Maple sugar and maple syrup	2099	Food Preparations, Not Elsewhere Classified (Major Group 20 Food and kindred products DIVISION-D Manufacturing)	111998	All Other Miscellaneous Crop Farming (111 Crop Production Sector)
6	5808101000	Braids, in the piece	2241	Narrow Fabric and Other Smallwares Mills: Cotton, Wool, Silk, and Manmade Fiber (Major Group 22 Textile mill products DIVISION-D Manufacturing)	315990	Apparel Accessories and Other Apparel (315-Apparel Manufacturing, 314-Textile Product Mills)

Figure B.1: Distances from Indirectly Linked Products

(a) Directly Linked Distances

$$\begin{bmatrix} 0 & v_{12} & NA & NA & NA & v_{16} \\ v_{21} & 0 & v_{23} & NA & NA & NA \\ NA & v_{32} & 0 & v_{34} & NA & NA \\ NA & NA & v_{43} & 0 & v_{45} & NA \\ NA & NA & NA & v_{54} & 0 & v_{56} \\ v_{61} & NA & NA & NA & v_{65} & 0 \end{bmatrix}$$

(b) Length 1 (c) Length 2



B "Filling in" Cost Bounds Algorithm

This section describes how to determine the upper and lower bounds for the distances between products that are indirectly linked. As shown in Figure B.1, we start with known upper and lower distance averages v_{ij} for pairs v_{12}, v_{23}, v_{34}, v_{45}, v_{56}, and v_{61}. These distances are based on direct co-production, with at one intermediary firm involved, assigning them a chain length of one.

Our next step involves recovering the distance bounds for pairs v_{ij} where there is no direct link but bounds can be inferred through a two-step chain via an intermediary product k, thus having a chain length of two (since it involves two intermediary firms). To recover these "chain length two" distances, we apply the triangle inequality. Specifically, we calculate the upper bound by finding the smallest maximum difference $|v_{ik} - v_{kj}|$ across all intermediaries k, and the lower bound by identifying the largest minimum sum $v_{ik} + v_{kj}$ across all k. These calculated bounds are then used to populate distances for indirect connections, as illustrated in Figure B.2.

We repeat this process with the newly populated distances, although now the upper and lower bounds of length 2 with correspond to length 3 separations of the original distances until $M - 1$ steps for M products are completed, which will result in a maxi min of lower bound distances and mini max of upper bound distances. This process is depicted in Figure B.3.

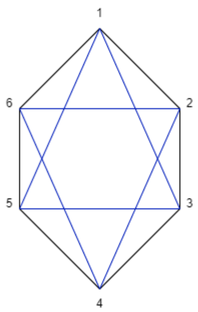
Once the matrices of upper and lower bound distances are populated as described, we compute the final distances d_{ij} as the average of the corresponding upper and lower v_{ij}.

Figure B.2: Distances from Indirectly Linked Products

(a) Indirectly Linked Distances

$$\left[\begin{tikzcd}
0 & v_{12} \arrow[r] & v_{13} & NA & v_{15} & v_{16} \arrow[l] \arrow[d] \\
v_{21} & 0 & v_{23} \arrow[u] \arrow[r] & v_{24} & NA & v_{26} \\
v_{31} & v_{32} \arrow[l] \arrow[d] & 0 & v_{34} \arrow[u] \arrow[r] & v_{35} & NA \\
NA & v_{42} & v_{43} \arrow[l] \arrow[d] & 0 & v_{45} \arrow[u] \arrow[r] & v_{46} \\
v_{51} & NA & v_{53} & v_{54} \arrow[l] \arrow[d] & 0 & v_{56} \\
v_{61} \arrow[u] \arrow[r] & v_{62} & NA & v_{64} & v_{65} & 0
\end{tikzcd}\right]$$

(b) Length 2



(c) Length 3

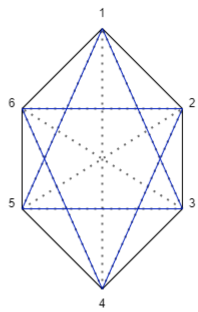
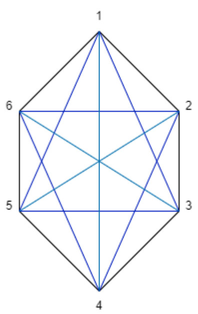


Figure B.3: Distances from Indirectly Linked Products

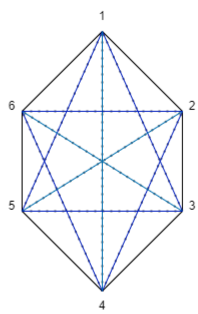
(a) Indirectly Linked Distances

$$\left[\begin{tikzcd}
0 & v_{12} & v_{13} \arrow[r] & v_{14} & v_{15} \arrow[d] & v_{16} \\
v_{21} & 0 & v_{23} & v_{24} \arrow[u] \arrow[r] & v_{25} & v_{26} \arrow[l] \arrow[d] \\
v_{31} & v_{32} & 0 & v_{34} & v_{35} \arrow[u] \arrow[r] & v_{36} \\
v_{41} & v_{42} \arrow[l] \arrow[d] & v_{43} & 0 & v_{45} & v_{46} \\
v_{51} \arrow[u] \arrow[r] & v_{52} & v_{53} \arrow[l] \arrow[d] & v_{54} & 0 & v_{56} \\
v_{61} & v_{62} \arrow[u] \arrow[r] & v_{63} & v_{64} & v_{65} & 0
\end{tikzcd}\right]$$

(b) Length 3



(c) Length 4



C Estimation

C.1 Distance Analysis

Stability of product space. Table C.1 provides the summary statistics of the growth rate in the distance between a firm and a product as $d_{\omega\nu t} - d_{\omega\nu t\ \ 1}$.

Table C.1: Changes in Firm-to-Product Distances: Summary Statistics (Values multiplied by 100)

	CES	DGKP
Average	-0.01	-0.05
Std. Dev.	10.26	23.67
5th P.	-17.07	-40.23
10th P.	-12.08	-29.47
25th P.	-5.73	-14.44
50th P.	-0.06	-0.18
75th P.	5.59	13.73
90th P.	11.44	27.64
95th P.	15.34	36.94

Shape and Characteristics of the Product Space. First, we focus on product-to-product distances, using the following regression model:

$$d_{\nu,\nu^0 t} = \text{Log Sales}_{\nu t} + \text{Log Sales}_{\nu^0 t} + \text{Co-production}_{\nu,\nu^0 t} + FE_{ct} + \epsilon_{\nu,\nu^0 t} \tag{C.1}$$

where Log Sales$_{\nu t}$ represents the total sales of multi-product firms for product ν. The term Co-production$_{\nu,\nu^0 t}$ quantifies the extent to which the two products are co-produced. This is measured either by the number of firms co-producing both products ν and ν^0, or by the proportion of firms that co-produce these products relative to the greater of the number of firms producing ν or ν^0.

Table C.2: Product-to-Product Distances

	Dependent Variable: Product-to-Product Distance					
	(CES)	(CES)	(CES)	(DGKP)	(DGKP)	(DGKP)
Log Sales ν	0.008***	0.003***	0.009***	0.009***	0.007***	0.009***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales ν^0	0.007***	0.002***	0.008***	0.009***	0.007***	0.009***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# Firms Co-producing ν, ν^0		0.233***			0.077***	
		(0.000)			(0.000)	
# Share Firms Co-producing ν, ν^0			0.715***			0.243***
			(0.001)			(0.001)
Cluster-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.06	0.41	0.41	0.07	0.09	0.09
# Obs.	1447074	1447074	1447074	1447074	1447074	1447074

Results from OLS estimation of (C.1). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.2 reveals that products with higher sales are generally more isolated than those with smaller sales. This observation is supported by the positive and statistically significant coefficient for Log Sales across all specifications. Contrary to expectations, the coefficient on co-production is also positive and statistically significant. This suggests that products often produced together tend to be more distantly located. Our analysis below indicates that firms are typically closer to the products within their production scope. Therefore, the observed positive correlation between co-production and product distance implies that our algorithm tends to position firms centrally among the products they produce, with those outside their scope being more distant from the firm but relatively closer to the products within the scope.

Further investigation in Table C.3 explores the possibility of a non-linear relationship between co-production variables and product distances. We observe a hump-shaped relationship, suggesting that products are further apart when only a few firms co-produce them. However, as the number of co-producing firms increases beyond certain thresholds (14 for CES and 10 for DGKP), the distance between products starts to decrease. Products co-produced by more than 28 firms in the CES specification and 21 in the DGKP specification tend to be closer than those never co-produced together. These findings from Tables C.2 and C.3 illustrate that our approach yields insights markedly different from those derived solely from co-production patterns.

Table C.3: Product-to-Product Distances

	Dependent Variable: Product-to-Product Distance			
	(CES)	(CES)	(DGKP)	(DGKP)
Log Sales ν	0.002***	0.005***	0.007***	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales ν^0	0.002***	0.004***	0.007***	0.007***
	(0.000)	(0.000)	(0.000)	(0.000)
Co-production (Number)	0.301***		0.104***	
	(0.000)		(0.001)	
Squared Co-production (Number)	-0.011***		-0.005***	
	(0.000)		(0.000)	
Co-production (Share)		2.018***		0.881***
		(0.002)		(0.004)
Squared Co-production (Share)		-1.643***		-0.805***
		(0.002)		(0.004)
Cluster-Time FE	Yes	Yes	Yes	Yes
R^2	0.48	0.56	0.10	0.11
# Obs.	1447074	1447074	1447074	1447074

Results from OLS estimation of (C.1). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

We extend our analysis to firm-to-product distances using the following regression model:

$$d_{\omega,\nu^0 t} = \text{Log Sales}_{\nu t} + \text{Log Sales}_{\omega t} + \text{Dummy for Produced}_{\omega,\nu^0 t} + FE_{ct} + \epsilon_{\omega,\nu^0 t} \qquad \text{(C.2)}$$

where Log $\text{Sales}_{\omega t}$ represents the total sales of firm ω and Dummy for $\text{Produced}_{\omega,\nu^0 t}$ is a binary variable that takes the value of one if firm ω manufactures product ν in year t.

The findings, presented in Table C.4, reveal some intriguing patterns. In the CES specification, firms are generally further from products with larger sales, whereas in the DGKP specification, the opposite trend is observed - firms are closer to products with larger sales. Moreover, there is a noticeable decrease in the distance of firms to products as the total sales of the firms increase. Additionally, firms are consistently closer to the products they produce, underlining a strong link between firm production profiles and their proximity to specific products in the product space.

Table C.4: Product-to-Product Distances

	Dependent Variable: Firm-to-Product Distance					
	(CES)	(CES)	(CES)	(DGKP)	(DGKP)	(DGKP)
Log Sales Product	0.002***	0.002***	0.002***	-0.001***	-0.001***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Sales Firm		-0.002***	-0.001***		-0.001***	-0.000***
		(0.000)	(0.000)		(0.000)	(0.000)
Production Dummy			-0.063***			-0.176***
			(0.000)			(0.001)
Cluster-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Product FE	0.11	0.11	0.13	0.10	0.10	0.13
Firm FE	1394297	1394297	1394297	1394297	1394297	1394297

Results from OLS estimation of (C.2). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

C.2 Product Drops

In this section, we quantify the ability of the product space to predict product drops. As the measure of distance within a firm is based on marginal costs that we observe, the estimation of the full product space is not necessary. As a result, we consider this exercise as a sanity check for our measure of marginal costs. For each firm, we select the products that the firm produces in its first year in the data. For instance, if a firm enters the dataset in 2000, the sample of products we consider are the products in the cluster of the firm that the firm produces in 2000. For each product in the years after entry, we compute a production indicator $\text{Drop}_{\omega,\nu ct}$ for product ν, firm ω, cluster c, and year t, which equals 1 if the product is not produced (i.e., dropped) and zero otherwise.

We estimate the following regression:

$$\text{Drop}_{\omega,\nu ct} = \beta Rank_{\omega,\nu ct\ \ 1} + a_{\omega t} + b_{\omega t} \tag{C.3}$$

where $Rank_{\omega,\nu ct\ \ 1}$ is the rank of produced products based on the distance of the product from the firm $d_{\omega,\nu ct}$, so that for the farthest product the product rank in the previous year equals one.

Results are shown in Table C.5 and C.6. Products that are farther from the firm are more likely to be dropped, as the coefficient on the lagged rank is negative and statistically significant in each specification. As shown for the case of product introduction, using a discrete classification cannot shed light on which products, within a certain aggregation of codes, are more likely to get dropped, while our measure based on marginal cost is able to distinguish between these produts.

Table C.5: Product Drop and Product Rankings

	Dependent Variable: Dummy=1 for Product Drop		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 farthest)	-0.034***	-0.062***	-0.006
	(0.009)	(0.007)	(0.007)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.60	0.60	0.60
# Obs.	22642	22642	22642

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.6: Product Drop and Product Rankings

	Dependent Variable: Dummy=1 for Product Drop	
	(CES)	(DGKP)
Lagged Rank (=1 farthest)	-0.023**	-0.056***
	(0.009)	(0.007)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.60	0.60
# Obs.	22642	22642

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

C.3 Export Shocks and Product Adoption

Table C.7: Export Shocks - IV = Autor et al. (2013) Countries

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	47.600	546.423***	25.540	101.694**	3.720	18.018
	(70.531)	(166.403)	(18.852)	(40.829)	(5.525)	(11.856)
(Lagged Log Distance)X(Log Exports)	-10.772*	-54.372***	-3.429**	-9.974***	-0.561	-1.806*
	(6.050)	(14.496)	(1.596)	(3.498)	(0.472)	(1.028)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782
F-Stat		1462.33		1810.90		1838.72

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Exports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k \in K}$ Exports$_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table C.8: Export Shocks - IV = EU Countries

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	47.600	509.799***	25.540	116.219***	3.720	14.576
	(70.531)	(149.486)	(18.852)	(37.006)	(5.525)	(10.952)
(Lagged Log Distance)X(Log Exports)	-10.772*	-51.171***	-3.429**	-11.222***	-0.561	-1.506
	(6.050)	(13.012)	(1.596)	(3.168)	(0.472)	(0.949)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782
F-Stat		1910.53		2346.02		2261.49

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log $\text{Exports}_{\nu t}$ Log $\text{Dist}_{\omega,\nu ct-1}$ with Log Exports $\text{IV}_{\nu t}$ Log $\text{Dist}_{\omega,\nu ct-1}$ where Log Exports $\text{IV}_{\nu t}$ = Log $\sum_{k\in K} \text{Exports}_{k\nu t}$ where K denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

C.4 First Stage Regressions

Table C.9: Export Shocks - IV = All Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.781***	0.792***	0.769***
	(0.002)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13782	13782	13782
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports $\text{IV}_{\nu t}$ = Log $\sum_{k\neq DNK} \text{Exports}_{k\nu t}$.

Table C.10: Export Shocks - IV = Autor et al. (2013) Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.781***	0.792***	0.769***
	(0.002)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13782	13782	13782
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports $\text{IV}_{\nu t}$ = Log $\sum_{k\in K} \text{Exports}_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table C.11: Export Shocks - IV = EU Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.781***	0.792***	0.769***
	(0.002)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13782	13782	13782
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports $\text{IV}_{\nu t} = \text{Log} \sum_{k \in K} \text{Exports}_{k\nu t}$ where K denotes denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

C.5 Import Shocks and Product Adoption

In this section, our focus is on examining the impact of an import shock on product adoption while considering its dependency on the proximity to the firm. Specifically, we employ a regression model, mirroring the export equation of the main text, to investigate this relationship:

$$\begin{aligned}\text{Intro}_{\omega,\nu ct} =& \beta_1 \text{Log Distance}_{\omega,\nu ct\ \ 1} + \beta_2 \text{Log Imports}_{\nu t} \times \text{Log Dist}_{\omega,\nu ct\ \ 1} + FE_{\nu t} \\ &+ FE_{\omega t} + FE_{\omega,\nu} + \epsilon_{\omega,\nu ct} \end{aligned} \tag{C.4}$$

where Log $\text{Imports}_{\nu t}$ represents the log of the total imports to Denmark of product ν in year t. All remaining variables and fixed effects are identical to those above for exports.

We instrument Log $\text{Imports}_{\nu t} \times$Log $\text{Dist}_{\omega,\nu ct\ \ 1}$ using Log Imports $\text{IV}_{\nu t} \times$Log $\text{Dist}_{\omega,\nu ct\ \ 1}$. We follow Autor et al. (2013) and use the imports of similar countries to instrument for Danish imports. In particular, we compute Log Imports $\text{IV}_{\nu t} = \text{Log} \sum_{k2K} \text{Imports}_{k\nu t}$ where K denotes two different sets of countries: the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA, and a set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden. The export data, sourced from BACI, is reported at the 6-digit level, while our products are defined more granularly at the 8-digit level.

The results of Tables C.12 and C.13 reveal a tendency for import shocks to exert a diminished impact on product adoptions when products are situated at a greater distance from firms. This observation aligns with findings from the export-level regressions. Corresponding first stage regressions can be examined in Tables C.14 and C.15.

Table C.12: Import Shocks - IV = Autor et al. (2013) Countries

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	67.253	300.371**	59.000***	86.377***	5.533	0.248
	(82.019)	(125.573)	(22.004)	(33.487)	(6.398)	(10.315)
(Lagged Log Distance)X(Log Imports)	-12.769*	-33.350***	-6.401***	-8.790***	-0.732	-0.268
	(7.141)	(11.021)	(1.898)	(2.908)	(0.554)	(0.901)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13719	13719	13719	13719	13719	13719
F-Stat		4926.69		5046.06		4131.44

Results from OLS estimation of (C.4). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Imports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Imports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Imports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Imports$_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table C.13: Import Shocks - IV = EU Countries

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	67.253	362.882***	59.000***	112.053***	5.533	6.881
	(82.019)	(131.994)	(22.004)	(34.458)	(6.398)	(10.563)
(Lagged Log Distance)X(Log Imports)	-12.769*	-38.869***	-6.401***	-11.031***	-0.732	-0.850
	(7.141)	(11.591)	(1.898)	(2.993)	(0.554)	(0.923)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13719	13719	13719	13719	13719	13719
F-Stat		4169.33		4577.43		3828.33

Results from OLS estimation of (C.4). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Imports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Imports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Imports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Exports$_{k\nu t}$ where K denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

Table C.14: Import Shocks - IV = Autor et al. (2013) Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Imports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Imports) IV	0.834***	0.842***	0.820***
	(0.001)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13719	13719	13719
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (C.4). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Imports IV$_{\nu t}$ = Log $\sum_{k\in K}$ Imports$_{k\nu t}$ where K denotes the same set of countries as Autor et al. (2013): Australia, Finland, Germany, Japan, New Zealand, Spain, Switzerland, and USA.

Table C.15: Import Shocks - IV = EU Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Imports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Imports) IV	0.834***	0.842***	0.820***
	(0.001)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13719	13719	13719
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (C.4). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Imports $\text{IV}_{\nu t}$ = Log $\sum_{k \in K}$ $\text{Imports}_{k\nu t}$ where K denotes denotes the set of EU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, and Sweden.

C.6 Marginal Costs in Levels

In this section, we replicate the baseline results presented in the main text. This replication uses the concept of marginal costs in level as a metric to quantify the distance between a firm and its products within our algorithm. Table C.16 provides summary statistics for the distances measured from product-to-product and firm-to-product. Tables C.17 and C.18 detail the correlation between the likelihood of product introduction and the rankings based on the firm's product distance. Tables C.19 and C.20 explore the relationship between the probability of introducing a product and its lagged distance to the firm. Table C.21 examines the impact of export demand shocks on product introduction, with the first stage regression detailed in Table C.22. Furthermore, Tables C.23, C.24, and C.25 investigate how Revenue Potential and Competition Potential influence the growth rates of sales, scope, and the Theil index of sales dispersion within a firm. Finally, Tables C.26 and C.27 present the outcomes of our counterfactual scenario, in which all firms introduce every variety.

Table C.16: Estimated Distances: Summary Statistics (2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.87	2.74	0.53	1.39
Std. Dev.	0.49	2.16	0.22	0.71
5th Perc.	0.21	0.52	0.22	0.38
10th Perc.	0.30	0.79	0.27	0.59
25th Perc.	0.48	1.00	0.37	0.91
50th Perc.	0.80	1.92	0.50	1.31
75th Perc.	1.11	3.99	0.65	1.76
90th Perc.	1.53	5.93	0.81	2.19
95th Perc.	1.75	6.91	0.91	2.51

Table C.17: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.005***	-0.007***	-0.008***
	(0.001)	(0.001)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.07
# Obs.	645290	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.18: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Rank (=1 closest)	-0.003***	-0.004***
	(0.001)	(0.001)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.19: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Distance	-0.698***	-0.596***	-4.672***
	(0.084)	(0.054)	(0.115)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.08
# Obs.	645290	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.20: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Distance	-0.280***	-0.285***
	(0.084)	(0.055)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.21: Export Shocks and Product Adoption

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	10.117	59.895***	1.462	11.970*	11.908	30.820
	(11.116)	(20.150)	(4.408)	(6.785)	(12.961)	(23.978)
(Lagged Log Distance)X(Log Exports)	-1.889*	-6.320***	-0.383	-1.271**	-1.553	-3.172
	(0.971)	(1.783)	(0.368)	(0.571)	(1.091)	(2.042)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782
F-Stat		2905.55		4908.25		2740.55

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Exports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$ is the total exports of all countries except Denmark.

Table C.22: Export Shocks - IV = All Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.768***	0.799***	0.772***
	(0.002)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13782	13782	13782
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports IV$_{\nu t}$ = Log $\sum_{k \neq DNK}$ Exports$_{k\nu t}$.

Table C.23: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	(CES)	(DGKP)
Lagged RP	0.542***	0.291***
	(0.075)	(0.046)
Lagged CP	-0.079*	-0.018
	(0.043)	(0.028)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.16
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.24: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	(CES)	(DGKP)
Lagged RP	0.028	0.061***
	(0.020)	(0.015)
Lagged CP	-0.135***	-0.062***
	(0.016)	(0.011)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.25: Firm Core Focus

	Dependent Variable: Theil Index (Core Focus)	
	(CES)	(DGKP)
Lagged RP	-0.058**	-0.049***
	(0.026)	(0.017)
Lagged CP	0.014	0.030**
	(0.017)	(0.012)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.77	0.77
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.26: Entry Potential by Sector (×100)

	CES		DGKP	
Cluster Number	Avg.	Std.	Avg.	Std.
Animal Products	72.5	21.5	77.3	22.2
Average	77.4	21.0	84.3	21.1
Chemical Products	71.7	23.7	80.0	27.5
Foodstuffs	77.8	22.3	82.3	24.5
Machinery and Electrical	70.3	16.1	75.9	21.7
Metals	82.1	24.6	89.9	18.6
Mineral products	70.2	20.1	78.0	25.3
Miscellaneous	81.3	19.4	92.0	15.0
Plastics and rubber	79.3	24.3	85.7	25.5
Stone and glass	77.5	22.8	82.8	23.6
Textiles	83.6	16.3	90.8	16.5
Transportation	85.3	19.2	92.4	15.8
Wood products	77.7	21.7	84.3	16.3
Average	**77.4**	**21.0**	**84.3**	**21.1**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index.

Table C.27: Entry Potential by Sector

	CES		DGKP	
	Avg.	Std.	Avg.	Std.
Green Products	0.82	0.22	0.87	0.25

Average and standard deviation of the EP for green products. We set $\sigma = 5$ in the calculation of the price index.

C.7 Inverse Hyperbolic Sine Marginal Costs

In this section, we replicate the baseline results presented in the main text. This replication uses the concept of marginal costs in level as a metric to quantify the distance between a firm and its products within our algorithm. Table C.28 provides summary statistics for the distances measured from product-to-product and firm-to-product. Tables C.29 and C.30 detail the correlation between the likelihood of product introduction and the rankings based on the firm's product distance. Tables C.31 and C.32 explore the relationship between the probability of introducing a product and its lagged distance to the firm. Table C.33 examines the impact of export demand shocks on product introduction, with the first stage regression detailed in Table C.34. Furthermore, Tables C.35, C.36, and C.37 investigate how Revenue Potential and Competition Potential influence the growth rates of sales, scope, and the Theil index of sales dispersion within a firm. Finally, Tables C.38 and C.39 present the outcomes of our counterfactual scenario, in which all firms introduce every variety.

Table C.28: Estimated Distances: Summary Statistics (2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.53	0.97	0.32	0.75
Std. Dev.	0.29	0.40	0.12	0.27
5th Perc.	0.15	0.36	0.15	0.26
10th Perc.	0.21	0.50	0.18	0.40
25th Perc.	0.31	0.72	0.24	0.59
50th Perc.	0.45	0.88	0.31	0.75
75th Perc.	0.68	1.16	0.39	0.90
90th Perc.	0.91	1.43	0.47	1.04
95th Perc.	1.06	1.66	0.52	1.13

Table C.29: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.004***	-0.005***	-0.008***
	(0.001)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.07
# Obs.	645290	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.30: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Rank (=1 closest)	-0.002***	-0.003***
	(0.001)	(0.000)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.31: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Distance	-2.718***	-1.160***	-1.688***
	(0.351)	(0.129)	(0.042)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.07	0.07	0.08
# Obs.	645290	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.32: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Distance	-1.230***	-0.548***
	(0.351)	(0.129)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.08	0.08
# Obs.	645290	645290

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table C.33: Export Shocks and Product Adoption

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	69.421	302.434***	15.288	70.766***	3.720	8.426
	(49.654)	(88.724)	(14.216)	(24.164)	(5.525)	(9.595)
(Lagged Log Distance)X(Log Exports)	-10.349**	-30.700***	-2.203*	-6.968***	-0.561	-0.971
	(4.256)	(7.704)	(1.203)	(2.065)	(0.472)	(0.830)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	13782	13782	13782	13782	13782	13782
F-Stat		3031.86		3540.91		3284.21

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log Exports$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ with Log Exports IV$_{\nu t}$ Log Dist$_{\omega,\nu ct-1}$ where Log Exports IV$_{\nu t}$ = Log $\sum_{k\neq DNK}$ Exports$_{k\nu t}$ is the total exports of all countries except Denmark.

Table C.34: Export Shocks - IV = All Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.782***	0.792***	0.769***
	(0.002)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	13782	13782	13782
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports IV$_{\nu t}$ = Log $\sum_{k\neq DNK}$ Exports$_{k\nu t}$.

Table C.35: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	(CES)	(DGKP)
Lagged RP	0.564***	0.328***
	(0.078)	(0.057)
Lagged CP	-0.075*	-0.018
	(0.046)	(0.030)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.16	0.16
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.36: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	(CES)	(DGKP)
Lagged RP	0.007	0.061***
	(0.021)	(0.018)
Lagged CP	-0.125***	-0.061***
	(0.017)	(0.011)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.13
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.37: Firm Core Focus

	Dependent Variable: Theil Index (Core Focus)	
	(CES)	(DGKP)
Lagged RP	-0.035	-0.062***
	(0.027)	(0.020)
Lagged CP	0.024	0.029**
	(0.017)	(0.012)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.77	0.77
# Obs.	8435	8435

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table C.38: Entry Potential by Sector (×100)

Cluster	CES Avg.	CES Std.	DGKP Avg.	DGKP Std.
Animal Products	0.72	0.21	0.75	0.23
Chemical Products	0.71	0.23	0.78	0.29
Foodstuffs	0.77	0.22	0.81	0.27
Machinery and Electrical	0.70	0.16	0.74	0.22
Metals	0.81	0.26	0.88	0.21
Mineral products	0.69	0.19	0.77	0.25
Miscellaneous	0.79	0.19	0.92	0.15
Plastics and rubber	0.79	0.24	0.84	0.27
Stone and glass	0.75	0.23	0.81	0.27
Textiles	0.83	0.16	0.92	0.16
Transportation	0.83	0.22	0.91	0.19
Wood products	0.74	0.25	0.79	0.20
Average	**0.76**	**0.21**	**0.83**	**0.23**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index.

Table C.39: Entry Potential by Sector

	CES Avg.	CES Std.	DGKP Avg.	DGKP Std.
Green Products	0.80	0.22	0.86	0.26

Average and standard deviation of the EP for green products. We set $\sigma = 5$ in the calculation of the price index.

D Larger Cluster Analysis

In this section, we replicate the baseline results presented in the main text, using a more aggregate definition of sectors. Specifically, we consider six sectors: Animals/Vegetables/Food (CN 2-digit 01-24), Minerals/Chemicals/Plastics (25-40), Textiles/Footwear (41-43, 50-67), Stone/Metals (68-83), Machinery/Transportation (84-89), and Miscellaneous (44-49, 90-97). This replication uses the concept of marginal costs in logs as a metric to quantify the distance between a firm and its products within our algorithm. Tables D.1 and D.2 present the descriptive statistics for the cluster characteristics. Table D.3 provides summary statistics for the distances measured from product-to-product and firm-to-product. Notice that relative to our baseline results, the average distances increase by little, with the larger increases concentrated in the higher percentiles.

Tables D.4 and D.5 detail the correlation between the likelihood of product introduction and the rankings based on the firm's product distance. Tables D.6 and D.7 explore the relationship between the probability of introducing a product and its lagged distance to the firm. Table D.8 examines the impact of export demand shocks on product introduction, with the first stage regression detailed in Table D.9. Furthermore, Tables D.10, D.11, and D.12 investigate how Revenue Potential and Competition Potential influence the growth rates of sales, scope, and the Theil index of sales dispersion within a firm. Finally, Tables D.13 and D.14 present the outcomes of our counterfactual scenario, in which all firms introduce every variety.

Table D.1: Cluster Descriptive Statistics

	Number of Products					Number of Firms				
Year	Avg.	Std.	Med.	25P.	75P.	Avg.	Std.	Med.	25P.	75P.
2000	207	141	191	84	260	134	64	111	90	196
2001	211	126	189	113	258	150	66	129	115	190
2002	208	141	174	91	260	143	72	129	83	190
2003	200	137	171	85	247	135	68	121	77	183
2004	205	123	175	118	250	144	56	133	128	182
2005	205	124	166	117	249	141	63	127	120	194
2006	211	128	174	124	258	136	60	130	105	192
2007	160	102	128	95	203	87	33	88	78	106
2008	163	107	130	111	208	84	34	88	79	93
2009	163	110	136	108	209	86	37	88	78	100
2010	173	97	143	123	203	100	41	111	87	131
2011	190	106	156	133	210	108	48	129	91	140
2012	195	103	162	135	225	100	40	112	82	133
2013	205	104	169	154	239	104	43	118	85	129
2014	212	108	167	166	267	105	44	117	94	129
2015	214	114	171	156	286	106	43	115	95	139
2016	208	120	160	154	288	100	46	109	80	140
2017	217	119	171	150	290	106	45	115	107	137
2018	219	123	158	151	292	107	45	118	100	140
Average	**198**	**118**	**162**	**125**	**247**	**114**	**50**	**115**	**93**	**150**

In each year, we compute average (Avg.), standard deviation (Std.), median (Med.), and 25th and 75th percentiles (25P. and 75P.) of the number of products (first four colums) and number of firms (last four columns) across clusters. In each year, there are 6 clusters (for 6 sectors defined as groups of CN 2-digit codes). The last row (Average) reports the average of the statistics across years.

Table D.2: Cluster Descriptive Statistics

	Number of Products		Number of Firms	
Cluster	Avg.	Std.	Avg.	Std.
Animals/Vegetables/Food	407	36	158	25
Machinery/Transportation	145	17	109	12
Minerals/Chemicals/Plastics	247	31	107	19
Miscellaneous	84	20	158	66
Stone/Metals	127	29	116	24
Textiles/Footwear	178	33	39	22

In each cluster, we compute average (Avg.) and standard deviation (Std.) of the number of products (first two colums) and number of firms (last two columns) across years.

Table D.3: Estimated Distances: Summary Statistics (2000)

	Product-to-Product		**Firm-to-Product**	
	CES	DGKP	CES	DGKP
Average	0.39	0.82	0.25	0.65
Std. Dev.	0.20	0.38	0.09	0.23
5th Perc.	0.12	0.32	0.11	0.29
10th Perc.	0.16	0.41	0.14	0.37
25th Perc.	0.24	0.59	0.18	0.50
50th Perc.	0.35	0.75	0.24	0.62
75th Perc.	0.48	0.99	0.29	0.77
90th Perc.	0.69	1.23	0.35	0.94
95th Perc.	0.74	1.42	0.40	1.05

Table D.4: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Rank (=1 closest)	-0.004***	-0.004***	-0.004***
	(0.000)	(0.000)	(0.000)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.05	0.05	0.05
# Obs.	1332970	1332970	1332970

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table D.5: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Rank (=1 closest)	-0.002***	-0.002***
	(0.000)	(0.000)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.06	0.06
# Obs.	1332970	1332970

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table D.6: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction		
	(CES)	(DGKP)	(CN)
Lagged Distance	-6.166***	-2.510***	-2.555***
	(0.419)	(0.153)	(0.045)
Product-Time FE	Yes	Yes	Yes
Firm-Time FE	Yes	Yes	Yes
R^2	0.05	0.05	0.05
# Obs.	1332970	1332970	1332970

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table D.7: Product Introduction and Product Rankings

	Dependent Variable: Dummy=1 for Product Introduction	
	(CES)	(DGKP)
Lagged Distance	-2.886***	-1.196***
	(0.421)	(0.153)
Product-Time FE	Yes	Yes
Firm-Time FE	Yes	Yes
CN Distance FE	Yes	Yes
R^2	0.06	0.06
# Obs.	1332970	1332970

Results from OLS estimation of (5). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%.

Table D.8: Export Shocks and Product Adoption

	Dependent Variable: Dummy=1 for Product Introduction					
	(CES)	(CES)	(DGKP)	(DGKP)	(CN)	(CN)
Lagged Log Distance	-53.495	430.422	48.417	125.458**	6.846	17.724
	(116.748)	(266.203)	(31.292)	(55.870)	(8.426)	(16.099)
(Lagged Log Distance)X(Log Exports)	-6.754	-48.850**	-5.755**	-12.347***	-1.022	-1.963
	(9.985)	(23.082)	(2.639)	(4.759)	(0.716)	(1.387)
Firm-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# Obs.	16266	16266	16266	16266	16266	16266
F-Stat		1882.60		3624.34		2964.54

Results from OLS estimation of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. We instrument Log $\text{Exports}_{\nu t}$ Log $\text{Dist}_{\omega,\nu ct-1}$ with Log Exports $\text{IV}_{\nu t}$ Log $\text{Dist}_{\omega,\nu ct-1}$ where Log Exports $\text{IV}_{\nu t}$ = Log $\sum_{k\neq DNK}$ $\text{Exports}_{k\nu t}$ is the total exports of all countries except Denmark.

Table D.9: Export Shocks - IV = All Countries (First Stage)

	Dependent Variable: (Lagged Log Distance)X(Log Exports)		
	(CES)	(DGKP)	(CN)
(Lagged Log Distance)X(Log Exports) IV	0.788***	0.797***	0.777***
	(0.001)	(0.001)	(0.001)
Firm-Time FE	Yes	Yes	Yes
Firm-Product FE	Yes	Yes	Yes
Prod-Time FE	Yes	Yes	Yes
# Obs.	16266	16266	16266
R^2	1.00	1.00	1.00

Results from OLS estimation of the first stage of (7). Standard errors in parenthesis. ***: significant at 99%, ** at 95%, * at 90%. Log Exports $\text{IV}_{\nu t} = \text{Log} \sum_{k \neq DNK} \text{Exports}_{k\nu t}$.

Table D.10: Firm Sales Growth

	Dependent Variable: Growth Rate of Sales	
	(CES)	(DGKP)
Lagged RP	0.452***	0.248***
	(0.071)	(0.056)
Lagged CP	-0.058	0.004
	(0.043)	(0.028)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.14	0.14
# Obs.	9552	9552

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table D.11: Firm Scope Growth

	Dependent Variable: Growth Rate of Scope	
	(CES)	(DGKP)
Lagged RP	-0.006	0.024
	(0.023)	(0.019)
Lagged CP	-0.128***	-0.062***
	(0.015)	(0.010)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.12	0.12
# Obs.	9552	9552

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table D.12: Firm Core Focus

	Dependent Variable: Theil Index (Core Focus)	
	(CES)	(DGKP)
Lagged RP	-0.064**	-0.067***
	(0.028)	(0.022)
Lagged CP	0.040***	0.022**
	(0.014)	(0.011)
σ	5	5
Cluster-Time FE	Yes	Yes
Firm FE	Yes	Yes
R^2	0.77	0.77
# Obs.	9552	9552

Results from OLS estimation of (16). Clustered standard errors in parenthesis. Cluster: firm. ***: significant at 99%, ** at 95%, * at 90%.

Table D.13: Entry Potential by Sector ($\times$100)

	CES		DGKP	
Cluster	Avg.	Std.	Avg.	Std.
Animals/Vegetables/Food	0.63	0.20	0.70	0.29
Machinery/Transportation	0.69	0.22	0.77	0.30
Minerals/Chemicals/Plastics	0.63	0.31	0.72	0.29
Miscellaneous	0.81	0.18	0.91	0.18
Stone/Metals	0.77	0.25	0.83	0.28
Textiles/Footwear	0.70	0.15	0.74	0.23
Average	**0.71**	**0.22**	**0.78**	**0.26**

Average and standard deviation of the EP for each cluster. We set $\sigma = 5$ in the calculation of the price index.

Table D.14: Entry Potential by Sector

	CES		DGKP	
	Avg.	Std.	Avg.	Std.
Green Products	0.76	0.23	0.83	0.28

Average and standard deviation of the EP for green products. We set $\sigma = 5$ in the calculation of the price index.