The rivalrous nature of know-how / The extent of rivalry in human capital / How industry-related capabilities affect export possibilities

Andres Gomez-Lievano[a], Sid Ravinutala[b]

[a] *Harvard University*
[b] *Harvard Kennedy School*

Abstract

Economic development occurs at the extensive margin of the technological frontier. However, many questions are still open about how technologies appear and diffuse. In particular, there is a puzzle in that technologies are the result of ideas, ideas are non-rivalrous, yet technology does not diffuse as easy as one would expect. Here we explore this question focusing on a particular context. We study events of export diversification in Colombian cities in the period 2008-2014. We find that the capacity of a city to expand the set of goods it can export should better be thought as a function of the availability of people with the expertise to do it. The concept of *expertise*, in contrast to the notion of *idea*, means that technology has a rivalrous aspect to it.

Introduction

We present evidence of rivalry as it applies to know-how, and we measure the extent to which it hinders economic growth. Our approach emphasizes the central role of know-how in economic processes. Know-how is distributed across brains of workers, and collective know-how is developed and expanded as different workers with different fields of expertise work together to create and produce more than what they would produce alone. The ways in which workers coordinate their know-how determines the products they can produce. Hence, different production processes correspond to different configurations of “pieces” of know-how. In this view, production can be expanded if one can expand the ways to combine and recombine different pieces of expertise and know-how that are distributed in the population of workers in a geographical location.

Our view differs from the traditional view in one important aspect, regarding the fundamental distinction between human capital and ideas [see 1]. In the traditional view, human capital acts as the source of ideas, ideas combine with ideas which leads to more ideas, which results in a view that focuses on the flow of ideas. Instead, in our work we emphasize the collaboration between individuals in teams. The former cares about the stock of knowledge an average individual holds in a society, while we care about the complexity of teams and their collective know-how. It does not matter how little individuals know on average in a society, but whether they manage to coordinate their individual know-how to accomplish more than they would if they acted alone. As a consequence, we posit that the constraints on economic development are primarily not about the limits to the flow of information into people’s head, but about the limited availability in a society of a diverse pool of know-how. What, thus, is the extent of rivalry of human capital for economic growth?

Background

Technology has been identified in the literature as the major determinant of economic growth and development. Technology understood as the embodiment of ideas and knowledge was originally advanced by "new growth" theories of Romer [2], Romer [3] and Lucas Jr. [4]. This focus on human capital as the source of technological advancement placed cities in the focus of attention to understand economic growth. Three empirical facts stand out: (1) Highly educated people concentrate disproportionately in larger cities [5, 6, 7]. (2) The population of places with high levels of human capital grows faster than places with low levels [8, 9]. And (3), increases in human capital entail gains in regional productivity and innovation [8, 5].

More recently, the literature has shifted its attention to the notions of specificity and interdependence in human capital. Studies have shown that when workers (and firms) specialize, they create dependencies to other workers. Thus, higher productivities are attained not by knowledge spillovers, but by know-how complementarities. This is a crucial distinction. The former implies that technology is constrained by institutions that hinder the flow of non-rivalrous ideas. The latter implies that technology is inherently constrained as by the rivalrous aspect of people's expertise.

Nobody challenges the fact that ideas are the engine of technologically-driven economic growth. From our point of view, however, human capital is highly differentiated and specific, and the technology space is explored by combining the know-how residing in people's heads, not the free flow of ideas while people bump into each other.

Data

The data we will work comes essentially from three sources: Administrative records of the social security system in Colombia (the Spanish acronym is PILA, for *Integrated Report of Social Security Contributions*), which contains data about sectors with 4-digit ISIC industry codes (the ISIC Rev. 3 A.C. codes, which are adapted to Colombian economic activities) and the information of employment sizes at the level of firms; Colombian customs data (the Spanish acronym is DIAN, for *Dirección de Impuestos y Aduanas Nacionales*), which contains the information about exports with 4-digit HS product codes (the HS codes 1992 revision) at the level of firms; And data about working age populations at the level of municipalities (publicly from the Colombian statistical office DANE). All sources, after merged together, cover the years 2008-2014. These are the data that are already being visualized in different ways and at different levels of aggregation in www.DatlasColombia.com.

There are firms that appear in ADUANAS that do not appear in PILA. This is due to the fact that PILA has been previously processed and many observations have been dropped (for example, because of information incorrectly reported, like industry or municipality codes, among others).

The final dataset contains 8,524,309 total rows. In it, there are 2,519,960 unique firms, across 7 years, 1,123 municipality codes, 62 city codes. Cities in Colombia are defined as the set of 19 metropolitan areas and 43 municipalities with populations above 50,000. The 19 metropolitan areas consist of collections of two or more municipalities, where a municipality belongs to a metropolitan area if at least 10% of its population commute to any of the other municipalities within the area [see 10]. We end with 1,179 product codes, and 469 industry codes.

We will use this firm-level dataset for two purposes: First, we will analyze, describe and identify the main statistical patterns of the firms that export. And second, we will aggregate the quantities and collapse the information to the level of cities, to understand and learn about the export possibilities across all 62 cities in Colombia.

Petroleum and mineral products in Colombia are not apt for the type of analysis that follows, in the sense that our analysis emphazises knowledge-based economic activities in which skills are located in the places of production. These products are raw materials which represented more than 63% of total Colombian exports in 2014, and these are products that are very sensitive to fluctuations in price. Hence, movements in price strongly affect the total exports of Colombia and may hide other relevant productive capabilities. Thus, we drop “Coal” (code 2701), “Crude petroleum oils” (code 2709), “Refined petroleum oils” (code 2710), and “Petroleum gases” (code 2711).

After cleaning our dataset, we observe that in 2013 the total number of effective number of employees in the formal sector was 6.88 million workers (which is consistent with what the DatlasColombia site reports, 6.7 million).

Descriptives

There are 2,519,960 unique firms that appear between 2008 and 2014 in our dataset, 21,026 firms report at least one exported good in any of the years. Table **??** breaks up these numbers by year, and we report the number of firms per year that do, and do not, export.

[INCLUDE TABLE]

Some of these exporting firms do not report employees. This may happen because in the cleaning and process of the data, some firms are dropped from the PILA dataset when there are misreported variables. For example, from the 7,076 firms that did export in 2014, 479 did not report employees (6.7% of 2014 exporting firms). These, however, only represent 2.7% of total exports in 2014, and so they do not represent a significant problem for analysis.

When aggregating over the municipalities for firms that have operations in many places, we end up with five quantities of interest at the firm-year level: total exports (in US dollars, or USD), total effective employment size, total number of different 4-digit codes the firm exports in (we will refer to this as the *number of products*), the average nominal wage paid per worker (in Colombian pesos, or COP), and the industry code the firm reported. Table 1 shows the basic descriptive statistics for the year 2014, for firms that exported.

[INSERT TABLE]

Tabelle 1: Descriptive statistics of exporting firms in 2014.

As we will see below, these descriptives must be interpreted with a bit of caution. This is because these quantities have very big variances, are very skewed and heavy-tailed. Notice, for example, that for all the variables the standard deviation is larger than the mean (in other words, the “coefficient of variation” is larger than one), and that the mean is always larger than the median. This is evidence that arithmetic averages for these quantities may not represent the typical values of the typical the firms. Below we will analyze the distribution of these quantities more in detail.

Relations between exports, size, wages, and number of products

Now, it is reasonable to suspect that larger firms export more in value, but presumably also in the number of products. One of these relationships is shown in Figure 1. In it, we have controlled for year fixed effects, and is the relationship for all years between the logarithm of firm size and the logarithm of total exports per firm. According to an OLS fit, the average relation is

$$x(n) \approx x_0\, n^{\gamma}, \tag{1}$$

where x are exports and n is the effective number of employees, $\widehat{x_0} \approx 22{,}000$ and $\widehat{\gamma} \approx 0.5$. In simple words, a small firm of size 1 starts with a total of 22,000 dollars in yearly exports on average (actually, it was 24,715 in 2008 and went down to 19,816 in 2014), and this grows with the square root of the number of employees. Hence, a quadrupling of the number of employees will be associated with a doubling of its exports.

[INSERT FIGURE]

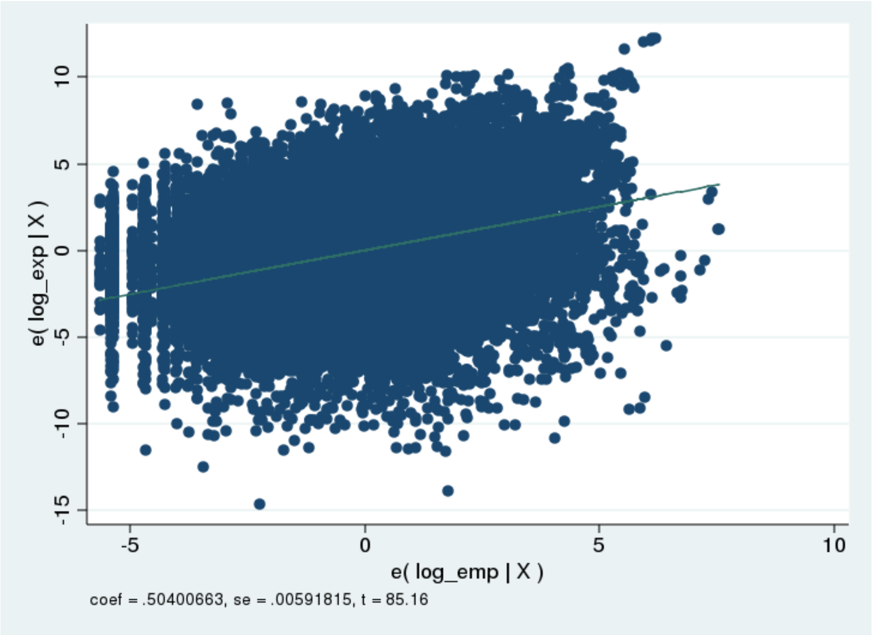

Figure 1: Larger firms export more. This is the scatterplot of the partial correlation of (log) total exports against (log) employment, controlling for year fixed-effects.

Firms that have high exports can grow in size, their growth in size allows them to use a larger pool of know-how, which should in turn increase the number of products they can export. As they grow, perhaps, firms also pay higher wages. Table 2 shows the pairwise correlations between these variables for firms in 2014.

Table 2: Pearson correlation between pair of variables across firms in 2014.

[INSERT TABLE]

In addition to the pearson pairwise correlations of Table 2, in Table 3 we show the pairwise *elasticities*. In the table, the row acts like the dependent variable and the column the independent variable, controlling for year fixed-effects.

[INSERT TABLE]

Table 3: Pairwise elasticities after controlling for year fixed-effects between rows (dependent variables) and columns (independent variables). Each element in the matrix is thus understood as the associated percent increase in the row variable if the column variable is increased by 1%.

From Table 2 we conclude that the pair of variables that correlate the most are number of products produced with total exports. And from Table 3 we observe that total exports change superlinearly *only* with the number of products (in contrast, total exports change sublinearly with employment size of the firm). Hence, the value of total exports in a firm is most responsive to the number of products the firm is able to export. This, of course, is just an association, and is difficult to assert which of the variables are causally determining which other. As we show below, however, there is evidence to believe that total exports are causally driven by the number of products the firm produces.

Below we show regressions of total exports against firm size and the number of products exported, and we also include the wages paid by firms as a control (see Table **??**). In the even columns of the table, we repeat the regression but we include industry fixed-effects.

[INSERT BIG TABLE REGRESSION]

From Table **??** we observe that the effect of firm size is partly taken away by the inclusion of wages and the number of products. We find again, however, that larger firms export more, and more products.

In the last column, controlling for everything else, a 1% increase in the number of products is associated with a 1.12% increase in total exports. The empirical piece that suggests that this may be causal (although it is not a rigorous argument) comes from the *shape* of the relationship between these two variables, shown in Figure 2. There, each dot is a firm in a year. In the right plot, we show the partial scatterplot correlation of the (log) of total exports agains size and against the (log) number of products, controlling for year and industry fixed effects, employment size and wage. As is clear, the shape of the scatter is triangular such that the bottom-right (many products and low exports) part of the plot has almost no firm. Starting from the bottom-left part of where the points are (i.e, firms that produce few products and have small total exports), one can see that increasing the number of products unavoidably leads to higher total exports, but not the other way: if a firm exports more (again, starting from the bottom-left) in value, that does not lead to more products. This is exactly the type of behavior expected from a causal relationship. It is a situation where there is a "*if* ***p*** *then* ***q***" type of statement, where "**p**" corresponds to the event "product diversification" and "**q**" stands for the event "increase in total exports".[1]

[INSERT FIGURE]

[1] Notice the use of the word "event". This is because "things" do not *cause* anything. Only events cause events.

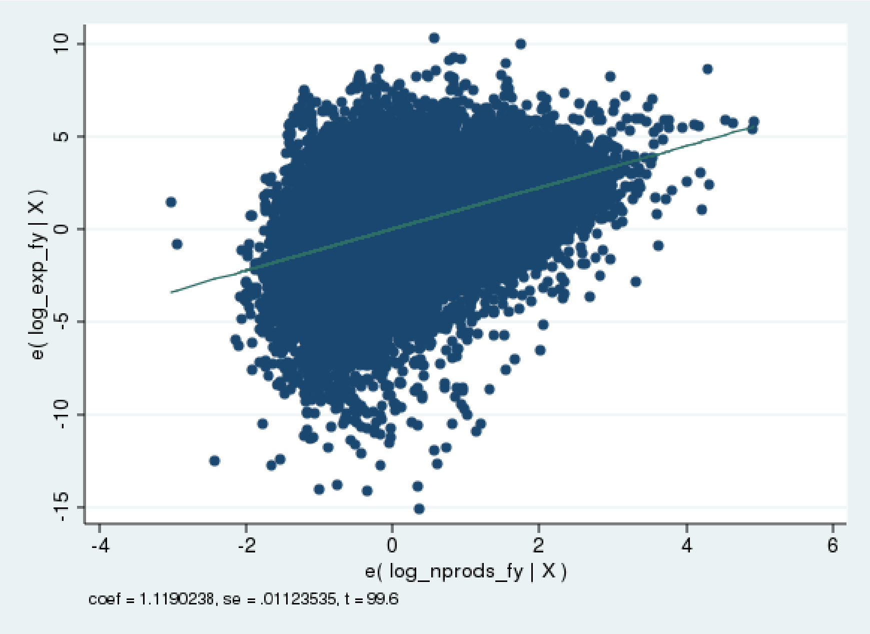

Figure 2: More diversified firms export more.

Presences of industries and exports in cities

The concept of "Location Quotient" or "Revealed Comparative Advantage"

Generally speaking, to identify presences, we typically use some measure of "representativeness", or "concentration", of an activity in a place. In urban studies, these are typically called "Location Quotients" (or LQs), and in the trade literature, "Revealed Comparative Advantage" (or RCAs). The general idea behind these measures is that it is a comparison, usually a ratio, between what is *actually* present and what is *expected* to be present, $RCA_{c,p} = X_{c,p}/\widehat{X}_{c,p}$. The expectation in the denominator requires one to have a "model of the world". The convention is to assume a very simple null model based on a law of proportionality. For example, a location c is expected to export product p in the same proportion as the product p is exported on average everywhere else. Thus, if the total exports of a location is X_c, and the average share of p is $\hat{s}_p$, then the expectation of how much c should export of p is $\widehat{X}_{c,p} = X_c \times \hat{s}_p$. The RCA, according to this null model, will be $RCA_{c,p} = X_{c,p}/(X_c\hat{s}_p)$.

It is important to note, however, that more sophisticated models can be constructed, which may increase our ability to identify the unexpected presence of economic activities in places.[2] For example, one may have a linear model that makes predictions $\widehat{X}_{c,p}$ based on a regression using some factors of interest. The RCA will thus be the ratio between what we actually observe and what our model predicts. If the quantity of interest $X_{c,p}$ is positive, then the logarithm of the RCAs gives us the residuals of our model. Or, conversely, if one has a model to explain $X_{c,p}$, the corresponding RCA is the exponential of the residuals of the estimated regression.

[2] See, for example, "Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities" by Bettencourt et al., PLoS ONE, 2010.

These ratios between "actual" over "expected" therefore provide us with dimensionless numbers that when larger than 1, there is a higher concentration of that activity than expected, or a lower concentration than expected if less than 1.

As implied before, a natural transformation is to take logarithms. In logarithmic scale, 0 is the point of reference above which you identify uncompetitive from competitive sectors (since $\log(1) = 0$). Since the statistical empirical distribution of RCAs is lognormal, it makes statistical sense to take logarithms. The reason for this is that RCA, being lognormally distributed, has non-negligible probability of being extremely large. Hence, one observes many RCA values that are less than 1, but some few that can be of the order of thousands or tens of thousands. Taking logarithms transforms such a heavy-tailed distributed variable into values normally distributed. There is a downside to this, however, because there are many RCAs which have the exact value of 0, and thus the logarithm returns $-\infty$. Thus, by taking logarithms one shrinks the extremely large positive values, but takes the 0s and throws them into minus infinity. To solve this problem we note that the *natural* logarithm can be expanded in the following power series:

$$\begin{aligned} \ln(x) &= \lim_{n\to\infty} \operatorname{approxlog}(x,n) \\ &= 2\lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{2k+1}\left(\frac{x-1}{x+1}\right)^{2k+1}. \end{aligned} \tag{2}$$

Thus, one can apply logarithms by applying this formula to a given finite order n. When applied to $x = 0$, the function "approxlog(x,n)" returns a finite negative number, and the larger the order n of the approximation, the more negative. We can do one additional transformation to "fix" that. At the end, these transformations will generate values that are (i) non-negative, (ii) normally distributed, and (iii) keep the same qualitative interpretation of the original RCAs with respect to the threshold of 1. The final transformation is to choose a large value $n >> 1$ subtract the function at 0 and then normalize by it:

$$modRCA_{c,p} = \frac{\operatorname{approxlog}(RCA_{c,p}, n) - \operatorname{approxlog}(0,n)}{-\operatorname{approxlog}(0,n)}. \tag{3}$$

We will apply this formula to both product and industry RCAs. Notice that we are translating and scaling a normally distributed random variable, and thus, the resulting variable "modRCA" (from *modified* RCA) is also normally distributed. When $RCA = 0$, the modified $modRCA = 0$ is also zero, and the same when $RCA = 1$, then $modRCA = 1$.

In most of the analysis below, we will try to be explicit about when we are using the untransformed RCA, or when we are using the modified RCA of eq. 3 (to a certain order n, although the convergence is very quick, so typically $n \approx 500$ is more than enough, although one has to make sure that n is such that $\operatorname{approxlog}(0,n) < \ln(x_{\min})$, where $x_{\min}$ is the minimum value in the data greater than zero). When the context demands it, we will explicitly distinguish between the real RCA and the modified RCA from eq. 3.

Regarding the economic interpretation of RCAs, a final clarification is worth mentioning. The name "Revealed Comparative Advantage" is strictly a misnomer because the measure does not capture in any way the actual advantage of doing things competitively, or efficiently. For instance, heavily subsidized exports will have a high RCA in spite of not being competitive in cost or resource-use terms. When RCAs or LQs are larger than 1, one typically says that there is a competitive advantage, but it only reveals that the place has a ”relatively large quantity” of X. Hence, we will use the expressions “relatively large”, “competitively”, and “highly concentrated” interchangeably.

RCAs assuming proportionality

For industries, our null model will be the proportion of formal employment in Colombia in a specific industry as a share of the working age population. Hence,

$$RCA_{c,i} = \frac{\frac{E_{c,i}}{W_c}}{\frac{\sum_{c'} E_{c',i}}{\sum_{c''} W_{c''}}}, \tag{4}$$

where $E_{c,i}$ is the total effective number of employees in city c assigned to industry i, W_c is the working age population in city c. Notice that eq. 4 will allow us to identify the places that have a high concentration of employment in an industry with respect to Colombian standards.

For exports, our null model will be based on the international standards of exports *per capita* for a specific product p. Hence,

$$RCA_{c,p} = \frac{X_{c,p}/P_c}{X_p^{\text{tot.}}/P^{\text{tot.}}}, \tag{5}$$

where $X_{c,p}$ is the exported value in city c of product p, P_c is the total population in city c, $X_p^{\text{tot.}}$ is the worldwide total exports of product p, and $P^{\text{tot.}}$ is the worldwide population. eq. 5 will allow us to identify the places that export products competitively with respect to international standards.

From an economic point of view, there is a trade-off for using one or the other formula for exports. If we use employment over working age population, we get rid off the fluctuations that come from the movements of prices that affect how much in value is exported (we do not want to claim that a city became good at exporting a product just simply because international prices for that product increased). However, the formula that uses exports per capita compared to the international exports per capita quantifies what exports really are about. Namely, the capacity to compete internationally in the production of a good. Hence, we will mainly be focusing on the RCA as given by eq. 5 for exports, but we will maintain the formula eq.4 when analyzing industries.

Regarding the distribution of *modRCA*’s that result from transforming eq. **??** using eq. 3, we show in Figure 3 the histograms associated with both industries and products, which make explicit the fact the logarithms of RCAs are approximately normally distributed. In a lognormal distribution, the highest point of the bell-shape density function marks the median (not the mean!), and we show in the figure the vertical gray line that divides the activities that are competitive from the uncompetitive. It is thus clear, again, that cities have presences of industries clustered around the value 1, but export products mostly below the threshold for being internationally competitive.

[INSERT FIGURE]

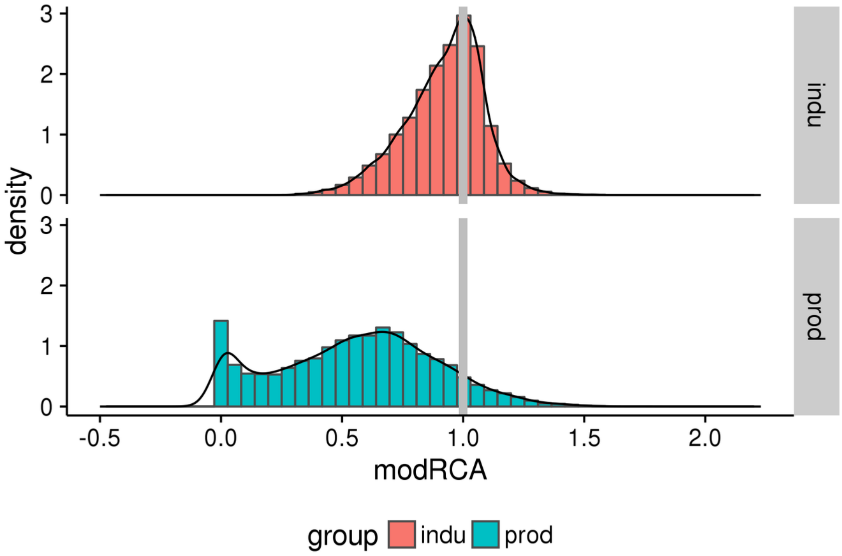

Figure 3: Histogram of *modRCA*'s for all cities and all industries (top plot) and all products except petroleums (bottom plot), which are a transformed version of *RCA*'s but rescaled such that they are now approximately normally distributed. The vertical gray line divides corresponds to *RCA=1*. Zeros not shown.

Matrices of presences

We begin simply by showing the general patterns of presences of industries and export products. The "scrabble theory of economic development" starts from the observation that the matrix of what places produce is *nested*. This observation is synonymous from saying that the matrix has a triangular pattern in it when rows and columns are properly ordered. The economic significance of this patterns is that it suggests that there is an underlying hidden state variable that is being accumulated as places get rich. Specifically, it suggests that places *add* capabilities to their productive processes, and therefore the number of things they produce increases. Hence, the conclusion that the process of economic development is one of accumulation and coordination of productive capabilities, and this process has as a consequence a pattern of diversification, not specialization.

Below in Figure 4 we show two different ways of visualizing the presence of economic activities. The columns are all the 465 industries (4-digit ISIC) together with all the $1,163$ products (4-digit HS). The rows are all the 62 cities in Colombia. The matrix on top is showing the continuous values of modRCAs, the matrix in the middle shows the discretization such that it is 1 (blue cells in the matrix) when $modRCA > 1$, and the matrix on the bottom shows the same discretized version but the columns have all been organized from least ubiquitous on the left to most ubiquitous on the right, regardless of whether it is an industry or a product.

The matrix of RCA's (top matrix in Figure 4) is the mean of the *modRCA* (see eq. 3) for each city and industry/product across all years (2008-2014) removing the two extreme values (i.e., removing the years for the smallest and largest RCA's of place in an activity). Each year, the RCA is discretized so that 0 is when $RCA < 1$ and 1 is when $RCA > 1$. That is what we call the figure "Binary presence". To illustrate representative presences across the seven years for which we have data, we show in Figure 4 the median binary presence across all 7 years. Since "7" is an odd number, we are essentially showing a 1 if the industry/product had an $RCA > 1$ for a *majority* of years, and 0 otherwise.

[INSERT FIGURE]

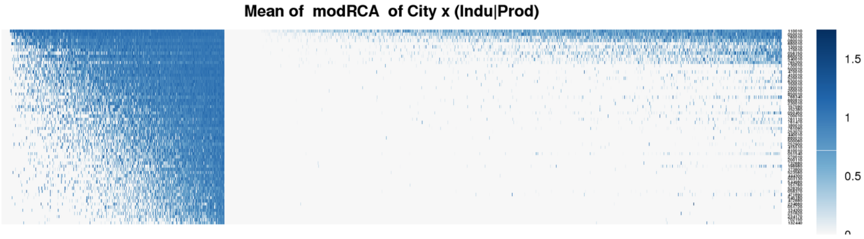

Figure 4: Matrix of average presences across years. All rows have been re-ordered such that the top-most city has the largest number of industry and product presences (together) and bottom-most city the fewest. It shows the *modRCA*'s (calculated at the city level within Colombia) for industries and products separately, and the columns are ordered so that the most ubiquitous are to the left.

In Figure 4 we observe the triangular pattern in industries and in products. We observe, not surprisingly, that the presence of products is more sparse than the presence of industries. Interestingly, there is a sharp cut in the products, whereby the left part of the matrix has only zeros revealing that cities are internationally competitive in very few products. For most products, the export RCA is below 1 in *all* cities.

Ubiquities per industry and per export

We observe from Figure 4 that some industries and some products seem to be present in relatively many cities, while some industries and some products seem to appear only in the largest cities (some do not appear anywhere). We say, accordingly, that some industries (or products) have high ubiquity, while others have low ubiquity. This notion of "ubiquity" is important because it indicates how difficult it is to promote a given economic activity, related either to employing people in a particular industry, or to exporting a particular product. To understand how the property of "ubiquity" changes across industries and products here we characterize industries and products by their overall presences across cities.

To quantify "ubiquity" the convention is to compute the sum of each of the columns of a matrix of binary presences, such as the matrices of the middle or bottom panels of Figure 4. Thus, by computing the so-called "colsum" of a presence matrix we get the vector of ubiquities. This method works well in general, but in our case it will hide important information. The reason is that industries have many presences in general while products have very few presences. Thus, we will hide the fact that products are, in fact, being produced and exported by several cities, only not at international standards given city-populations (see Figure **??**).Figure **??**). Hence, we will quantify the ubiquity of a product not as the integer count of all the cities in which it had $RCA > 1$, but, rather, as the sum of its *modRCA*. Recall that *modRCA* has transformed the original *RCA* such that it is now normally distributed, as opposed to lognormally distributed, but this transformation has maintained the mass of the distribution on either side of the value 1 unchanged. Thus, we define the ubiquity of industries and products (separately) as

$$\text{Ubiquity}_i = \sum_c modRCA_{c,i},$$
$$Ubiquity_p = \sum_c modRCA_{c,p}.$$

We would not want to add the unmodified RCAs, because the sum of such heavy-tailed distributed values will be dominated by the extreme values, as opposed to the common or more representative values. On the other extreme, we have decided not to use the binary presences, since it neglects presences of activities in places that, while not internationally competitive, are not zero. Equations **??** are a compromise between these two extremes.

Figure **??** plots the histograms of the ubiquities of industries and products. Interestingly, the distribution of ubiquities across industries appears to be relatively flat, such that industries with both very high and very low ubiquity (i.e., industries that are very common and very uncommon, respectively) are rare, but for most of the intermediate range between 0 and 62 (the minimum and maximum possible ubiquities) ubiquities are approximately uniformly distributed. In contrast, the bottom panel of Figure **??** shows that products have for the most part very low ubiquities. That means that most products are generally not exported, or exported in very low quantities.

[INSERT FIGURE]

Roughly speaking, there are 200 products that are basically not produced anywhere in Colombia, 250 exported in just one city, and other 200 products exported in two cities. That is why the median of the distribution of product ubiquity is approximately 2. There is an outlier, however, which is a product with ubiquity of 30, which is “Non-roasted coffee” (product code 0901).

Diversities of cities, with respect to industries and products

We observe from Figure 4 that rich cities have many industries and export relatively many products, while less developed cities have few industries and export few or no exports. Here we show a characterization of cities, in terms of how many industries they have, and how many products they export. In the same way and for the same reasons we mentioned regarding the ubiquities of industries and products, we will quantify the industry diversity and the product diversity of each city as:

$$\text{InduDiversity}_c = \sum_i modRCA_{c,i},$$
$$ProdDiversity_c = \sum_i modRCA_{c,p}.$$

This is a rough estimation of how many industries, or how many products exported, respectively, a city *c* has.

[INSERT FIGURE]

In Figure **??**, we show the histograms of the diversities. When looking at the number of industries a city could have, we conclude from this figure that there is a wide spread of values. Starting from cities that have less than 100 industries, to a few cities that have all of them (~ 465).

The median number of industries present in a city is 221. As opposed to this broad range of industry diversities, product diversities are very low. The median product diversity is 5. Yet, the values of product diversity display some extreme values. In particular, one can observe from Figure **??** that there are five cities that export a disproportionately large number of products. Bogota Met has a product diversity of 622, Medellin Met of 536, Cali Met of 377, Barranquilla Met of 296, Rionegro Met of 195, and Cartagena Met of 155. While exports are highly concentrated in the largest cities, Rionegro stands as an interesting exception suggesting that export capabilities are not a mechanical result of city size.

We suspect that having more industries leads to more products to export. In fact, this is the premise of this whole study. To see this relationship more generally, we show the scatter plot of industry diversity versus product diversity in Figure 5, where each dot is a city. We present industry diversity as a percentage of the total number of industries and product diversity as a percentage of total number of products. Thus, what we are plotting is the *share* of total industries versus the *share* of total number of products, and how they change across cities. All four panels have exactly the same information, all of them with the industry diversity share in the x-axis and product diversity share per city in the y-axis. However, in the different panels we show the axes with different linear and logarithmic scales, to reveal which of the following relationships describes the data the best: linear (top-left), logarithmic (top-right), exponential (bottom-left), or power-law (bottom-right). As a reference, we have included a dotted gray line which corresponds to the equation $y = x$.

[INSERT FIGURE] [h!]

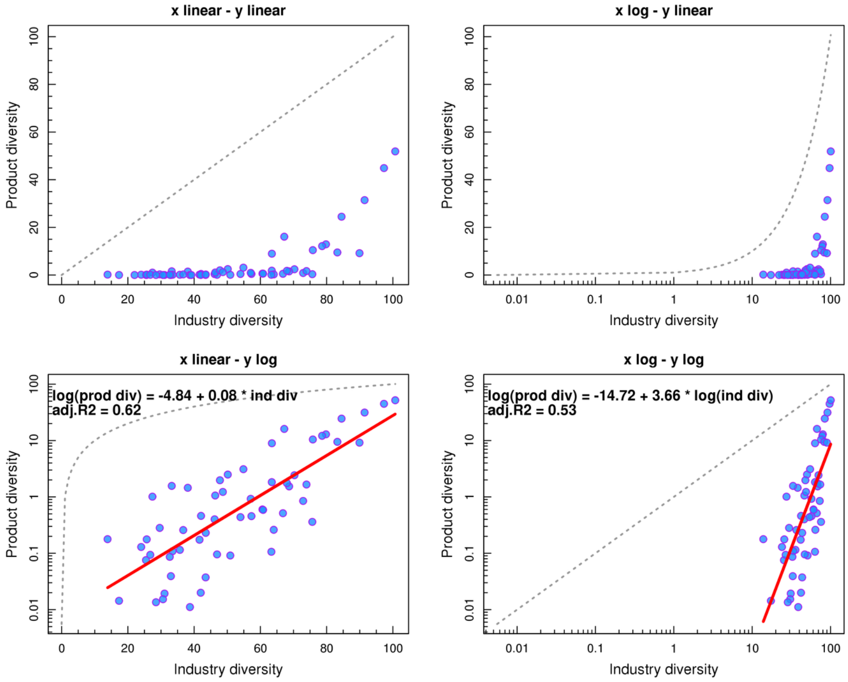

Figure 5: Scatter plot of diversities per city, where each panel has different scales of the axes, to reveal whether there is a linear (top-left), logarithmic (top-right), exponential (bottom-left), or power-law (bottom-right) relationship. The gray dotted line represents the identity line (number of industries equal to number of products). Hence, the dots above the gray dotted line are cities that export more products with comparative advantage (with respect to other Colombian cities) than they have industries with comparative advantage.

It appears to be the case that the best relationship between industry diversity and product diversity is, either, (a) an exponential relationship

$$ProdDiversity \approx P_0 \ \mathrm{e}^{g \ InduDiversity},$$

where $\hat{g} \approx 0.082$ (with a standard error of 0.008) such that the addition of 1 percent more industries to a city's basket is associated with a 0.082% increase in the percentage of products; or is (b) a power-law,

$$ProdDiversity \approx P_0 \ InduDiversity^g,$$

such that increasing the number of industries by 1 percent is associated with an increase of $\hat{g} \approx 3.658$ percent the number of products (the standard error of the exponent is 0.45). Table **??** shows the results of these fitted regression models.

[INSERT TABLE]

In both models, the conclusion is qualitatively the same: adding industries has a dramatic effect on product diversity (with the traditional caveat that there is probably some reverse causality). In other words, products accumulate and concentrate very quickly as cities industrialize. If there is a causal connection between increasing the diversity of industrial employment and increasing the diversity of exported products in a city, we can expect from these results to see huge increases in exports with relatively modest industrialization.

As a final remark, the exponential relationship between industrial diversity and product diversity is consistent with a model in which industries are the ingredients that get combined in cities to produce and export products.[3] Suppose q is the probability that any product requires a specific industry as one of its ingredients. Suppose that some products are made of many ingredients while others just require few. Hence, the probability that an n-ingredient product exists is q^n. This evidences, based on a very simple probabilistic argument, that products that require many ingredients (large n) will be difficult to produce, and thus will be rare ($\lim_{n\to\infty} q^n = 0$). Hence, setting up this model in this simple way allows us to refer to n as the "complexity" of an n-ingredient product.[4]

Suppose now that a city has an industrial diversity D_I. With that many industries, there are

n$possiblecombinations(i.e.,products)of$n$ingredients(i.e.,industries).However,onlyafraction$q^n will exist. Thus, there will be on average

n q^n products of complexity n in a city with D_I industries. Hence, on average, the product diversity a city will have, D_P is

D$_P = \sum_{n=0}^{D_I}$
n q^n,
$= (1+q)^{D_I}$.

[3] See `https://arxiv.org/ftp/arxiv/papers/1601/1601.05012.pdf`.

[4] Note that this is a definition of product complexity "from first principles". As such, it differs from that used in Datlas, which is a statistical estimate. Furthermore, it is important to note that this statistical estimation of complexity is computed with information solely based on products, not of industries.

Assuming the probability q of requiring a industry as an ingredient is very small $q \ll 1$,

$$\mathrm{D}_P = \mathrm{e}^{D_I \ln(1+q)},$$
$$\approx \mathrm{e}^{q\ D_I}.$$

We see, thus, that product diversity in eq. b is an exponential function of the industry diversity, which is approximately what we observe empirically.

To make the connection between our fitted regression, some further manipulations are needed. If N_I and N_P are the total number of possible industries and products in a given classification, respectively, the above equation can be re-written in terms of shares as

$$\mathrm{M}_P \approx$$
$$\mathrm{M}_P \mathrm{e}^{(q\ M_I)\frac{D_I}{M_I}},$$

or more compactly

$$\mathrm{s}_P \approx$$
$$\mathrm{M}_P \exp\{($$
$$100 s_I,$$

where s_I and s_P are the share of industries, and the share of products, expressed as percentages, exactly as in the previous results (Figure **??** and Table **??**). If we believe this model, the estimates from Table **??** tell us that the probability that a product requires an industry is $\hat{q} \approx 0.018$.

The rivalrous effect of know-how in urban export diversification

Let us recall that we are interested in the mapping between the collection of pieces of know-how that a city has and what it can produce and export. Given the exact knowledge about that mapping the policy maker would be able to tell which products a city would be able to export given its existing know-how. Ideally, then, the strategy would be to infer a mapping between know-how and production, quantify the body of know-how of a place, and use the mapping to predict what the place could produce. We have said, however, that the individual pieces of know-how are difficult to measure and quantify, and as a consequence, inferring the mapping is not trivial.

The alternative strategy to solve these issues is instead using what a region produces as a guide itself to make educated guesses about what it could produce. This is done by constructing a so-called “similarity matrix” that quantifies how likely it is that a place exports a product given that it exports another product. A similarity matrix between products is what really defines the Product Space in Datlas (based on the International Atlas of Economic Complexity). Although less common, one can also create a similarity matrix between places.

These exercises can be understood to be part of the literature on “Collaborative Filtering”, or more generally, as “Recommender Systems”, in Machine Learning. This was popularized with the Netflix Challenge, in which the idea was to predict which movie would be a good recommendation for a given user. Formally, one has a matrix of users (as rows) and the movies (as columns) that users have watched. The idea, then, is to predict which zeros are the most likely to be filled in by the users. Notice that in our case we are recommending a product to a city. A popular collaborative filtering algorithm is nearest neighbors, which can be item-based or user-based. In the language of this literature, what we are trying to do here is item-based nearest neighbors collaborative filtering. Thus, the prediction is that a city will produce products which are similar to the ones it already produces.

The notion of the similarity matrices is crucial in understanding how regions diversify their economic activities. Based on different measures of similarity (between products and industries, and between products), we pursue the following main goal: to understand how exports grow and appear in places. For this, we construct measures of the potential a place has to export a product. We expect these measures to be predictive of production possibilities.

Trimming the set of products and industries

Before we carry out our analysis of similarities, we must deal with one final problem. A “similarity” between two economic activities (e.g., industrial employment, or exporting a product) has to be inferred from data, and is thus an estimate of an average relationship. As a consequence, we need a reasonably large sample size to establish such relationship.

One problem that arises here is that there are products that appear very rarely. If there is only one city in which a product is present, then there is nothing one can do to understand the requirements to produce it. Conversely, there are industries which are everywhere, which is also a problem. If there is an industry that is present in all cities, then it becomes uninformative about its effect on the presence or absence of products. Roughly speaking, to be able to have adequate measures that relate industries to exports, we need to remove the least ubiquitous products and the most ubiquitous industries.

Below we define the criteria we used to drop industries and products, and then we show the effects that dropping these have on our totals of employment, firms, and export values.

Criteria

We remove from our dataset any industry that satisfied *any of* the following four criteria in 2014:

- Has a ubiquity (eq. b) larger or equal than the top 95th percentile of industry ubiquities according to their $RCA_{c,i}$ of employment over working age population (eq. 4). This consists of 24 industries, which include industries 7499 ("Other business activities n.e.c.") and 4530 ("Building of civil engineering works"), both of which are present across many cities.

- Has a total size of people employed larger or equal than the other 95th percentile of industries in Colombia. This consists of 18 industries, which include industries like 0112 ("Cut flowers"), 1810 ("Manufacture of wearing apparel, except fur apparel") and 8050 ("Higher education"). Last two economic activities would be traditionally considered to be important for complex production processes, so it may seem unintuitive to drop them in our analysis. However, the criterion we are using here to drop them is reasonable given these already employ the most people, and in that sense they are probably easy to adopt in new places, and so they are unlikely to constitute a binding constraint for opening new export possibilities.

- Has a standard deviation $stddev_i$ of all its $modRCA_{c,i}$ across cities c smaller or equal than the bottom 5th percentile of the standard deviations of other industries. This consists of 24 industries, which include 9309 ("Other service activities n.e.c."), 8511 ("Hospital activities"), and 7491 ("Labour recruitment and provision of personnel"). In other words, service sectors that grow organically in every city regardless of the industrial structure (and are therefore uninformative for our purposes).

- Finally, since we also do not want industries which only appear in very few cities, we identify the industries that have ubiquities (eq. b) smaller or equal than 2 (according to their $\sum_c modRCA_{c,i}$). This consists of 12 industries, which include health and social activities such as veterinary activities, or animal husbandry.

Each criteria selects a set of industries, and we take the *union* of those sets. Under these criteria, the number of industries that will be dropped is 59 out of 468 (see Appendix **??** for the full table of dropped industries). In terms of their 2-digit categories, many of these dropped industries belong to wholesale trade, retail trade, and services sectors (e.g., accounting, legal services, domestic services, health services, repair of motor vehicles and motorcycles, etc.).

For exports, we remove any product that satisfied *any of* the following two criteria in 2014:

- Has a ubiquity (eq. b) smaller or equal than 2, according to its $modRCA_{c,p}$ (eq. 5). This consists of 542 products, which include products like 2844 ("Radioactive chemical elements"), 8526 ("Radar"), 8710 ("Tanks and other armored fighting vehicles"), and 3706 ("Motion-picture film").

- Has a total size of people employed smaller or equal than the other 10th percentile of the employment numbers associated with other products in Colombia. This consists of 105 products, which include products like 1204 ("Linseed"), 7903 ("Zinc powders"), 0101 ("Horses"), and 2940 ("Sugars, chemically pure, other than sucrose, lactose, maltose, glucose and fructose").

Under these criteria, the number of products that will be dropped is 546, out of 1175 (see Appendix **??** for the full table of dropped products). This is almost half of all products. All these products have very low ubiquities (due in part to the first criterion, obviously). The dropped product with the largest ubiquity is 7903 ("Zinc powders") with ubiquity 2.63 (recall that our definition of ubiquity according to eq. b allows non-integer ubiquities). Out of the 546 dropped products, 123 have a ubiquity of *strictly zero*.

[We briefly review the effects of this trimming of industries and products in the APPENDIX]

Mathematical definitions of Density

As a practitioner, one is typically interested in the following question: Do I have in my city the ingredients necessary to produce product p? If yes, how much of those ingredients do I have access to?

Below we develop four measures that try to answer these questions, by quantifying the intensity of the ingredients available to produce a product p in a city c. These different definitions are different, yet subtle, manipulations of the same basic equation, but these subtle differences actually lead to different predictions. We will show that one particular measure stands out as the best index to answer the question we start with in this section.

Suppose a city c "wants" to export good p. Our density measure, regarding a city c and a product p, should be something that can be interpreted as the *expected intensity of the ingredients city c has available that can contribute to the production of export good p.*

If we assume that each ingredient a has an additive effect on the ability to produce product p, the problem can be mathematically expressed as the product of two matrices:[5]

$$\underbrace{\mathrm{D}(t)}_{c\times p} = \underbrace{\mathscr{C}(t)}_{c\times a} \cdot \underbrace{\mathscr{P}(t)}_{a\times p}.$$

One the one hand, there is the matrix $\mathscr{C}$ of cities (as rows) and the amount they have of each ingredient a. On the other, one has the matrix $\mathscr{P}$, which lays out the products (as columns) and how much of each input a they each require in order to be produced. The element $[\mathrm{D}]_{c,p}$ is thus the potential of city c to produce product p, and is what we call the "density". We have explicitly written the time-dependence because all these matrices can change in time. However, to make the equations less cluttered, the time dimension will be dropped in what follows.

Equation b is the analytic basis behind our "density regressions". In practice, the equation leaves room for how to construct both matrices on the right-hand side, and on how to define what the ingredients a are. In what follows we will present four different interpretations of eq. b.

Density #1: $D^{(1)}_{c,p}$

In this first definition we will interpret eq. b in the following ways:

[5] The assumption of additivity is crucial to express the problem as the product of two matrices. For that reason, it is convenient to make that assumption. However, this assumption is probably not realistic. Hence, it is important to keep in the back of our minds that the reality is probably more close to a production function like a Leontief, where one has to have *all* ingredients to produce at least one unit of output.

[(a)]We will consider the industries to be the ingredients that drive exports, so i will denote the index of industries. $\mathscr{P}$ will be a $industry \times product$ matrix of weights based on a conditional probability. Specifically, the weights will be proportional to the probability that a worker is employed in industry i given she is employed in a firm that exports product p. The matrix $\mathscr{C}$ will be a relative intensity of employment in city c in industry i. Specifically, we will use $modRCA_{c,i}$.

Expressing explicitly all the elements that go into the construction of this first density, we have

$$\mathrm{D}_{c,p}^{(1)} = \textstyle\sum_i modRCA_{c,i} \; (\\ \textstyle\sum_{i'} \mathrm{Pr}_{(worker)}(i'|p),$$

where

$$\mathrm{Pr}_{(worker)}(i|p) = \\ \mathrm{E}_p/E_{tot}.$$

Density #2: $D_{c,p}^{(2)}$

In this second definition we will make a slight change to the first definition's way of estimating the conditional probability. The assumptions will be:

[(a)]We will consider the industries to be the ingredients that drive exports, so i will denote the index of industries. $\mathscr{P}$ will be a $industry \times product$ matrix of weights based on a conditional probability. Specifically, the weights will be proportional to the probability that an industry i is present in a city conditioned on the city already exporting product p. The matrix $\mathscr{C}$ will be a relative intensity of employment in city c in industry i. Specifically, we will use $modRCA_{c,i}$.

Expressing explicitly all the elements that go into the construction of this second density, we have

$$\mathrm{D}_{c,p}^{(2)} = \textstyle\sum_i modRCA_{c,i} \; (\\ \textstyle\sum_{i'} \mathrm{Pr}_{(city)}(i'|p),$$

where

$$\Pr{}_{(city)}(i|p) =$$
$$\mathrm{u}_p/N_c,$$

where N_c is the total number of cities, u_p is the ubiquity of product p, and $J_{p,i}$ is the number of cities in which industry i and product p were simultaneously present. In the statistical analyses below, we will construct J using Equation **??** but with the matrices of *modRCA*.

(Sometimes it is useful to express everything in terms of matrices and products of matrices. Let $X_{(c,i)}$ be the matrix of industry *modRCA*'s across cities, $M_{(c,i)}$ be the presences of industries in cities, and let $M_{(c,p)}$ be the presences of products across cities. Given this notation, the ubiquities of industries and products are

$$[\mathrm{I}]_i = \sum_c [M_{(c,i)}]_c$$
$$]_p = \sum_c [M_{(c,p)}]_c.$$

Given this, as well,

$$\{\mathrm{i—p}\}=$$
$$[\mathrm{P}]_p/N_c$$
$$= N_c \times \left[D^{-1}_{(p)} \cdot M^T_{(c,p)} \cdot M_{(c,i)} \right]_{p,i},$$

where $D_{(p)}$ is a matrix of zeros that has in the diagonal the ubiquities of the products.)

Density #3: $D^{(3)}_{c,p}$

The third definition that we will implement involves some additional opperations and some implicit matrix multiplications, but it can still be seen as a version of eq. b. The fundamental change is that the matrix $\mathscr{P}$ is now going to be interpreted as a similarity matrix between products. This similarity, however, will not be calculated based on the co-occurrence of products with products across cities. Instead, it will be based on a correlation measure between the vectors that define how products co-occur with industries. In this context, we make the following assumptions:

[(a)]We will still consider the industries to be the ingredients that drive exports (but as such, their appearance will be less explicit in the equations). $\mathscr{P}$ will be a *product* × *product* similarity matrix, based on a simple Pearson correlation between the rows of another *product* × *industry* matrix, meant to represent a sort of input-output matrix. This latter matrix will consist of normalized co-occurrences between industries and products across cities. The matrix $\mathscr{C}$ will be a relative intensity of export values in city c in product p. Specifically, we will use $modRCA_{c,p}$.

Expressing explicitly all the elements that go into the construction of this third density, we have

$$\mathrm{D}_{c,p}^{(3)} = \sum_{p' \neq p} modRCA_{c,p'} \; (\\ \sum_{p'' \neq p} \mathrm{cor}(\mathbf{p}'', \mathbf{p}),$$

where **p** are the rows of the matrix defined by the elements

-approxlog(0,500),

where $J_{p,i}$ is the co-occurrence, and u_p and u_i are the ubiquities, all three terms using *modRCA*'s. It is important to note that the interpretation of the last mathematical expression Equation b is simpler than it appears. It is simply a statement of whether product p and industry i co-occur in cities more frequently than what is expected.

Density #4: $D_{c,p}^{(4)}$

Finally, our fourth definition of density is practically identical to the third definition in Equation b. But to see the difference, note that in Equation b we are normalizing by the sum of the correlations, which means that we are taking an average of the $modRCA_{c,p'}$ of a city c across the products p', weighted by how correlated p' are to p. In our fourth definition, we will instead normalize by the sum of the $modRCA_{c,p}$, meaning that our density will be the average of the correlations between p and the rest of products p', weighted by how present p' are in city c. Thus, the assumptions are almost unchanged:

[(a)]We will still consider the industries to be the ingredients that drive exports (but as such, their appearance will be less explicit in the equations). $\mathscr{P}$ will be a *product* × *product* similarity matrix, based on a simple Pearson correlation between the rows of another *product* × *industry* matrix, meant to represent a sort of input-output matrix. This latter matrix will consist of normalized co-occurrences between industries and products across cities. The matrix $\mathscr{C}$ will be a relative intensity of export values in city c in product p. Specifically, we will use $modRCA_{c,p}$.

Expressing explicitly all the elements that go into the construction of this fourth density, we have

$$\mathrm{D}_{c,p}^{(4)} = \sum_{p' \neq p} \\ \sum_{p'' \neq p} modRCA_{c,p''} \; \mathrm{cor}(\mathbf{p}', \mathbf{p}),$$

where **p** are the rows of the matrix defined by the elements

-approxlog(0,500),

where $J_{p,i}$ is the co-occurrence, and u_p and u_i are the ubiquities, all three terms using *modRCA*'s. Exactly as in our third density above, Equation b is simply a statement of whether product p and industry i co-occur in cities more frequently than what is expected.

Empirical results

Before we present our empirical results, it is important to state in words the interpretation of each of our densities, Equations **??**, when making a reference to a specific city c and a specific product p:

$_{c,p}$:] Weighted average of the concentration of employment in our city c across all indutries $i \in \{1,2,\ldots\}$, with weights $w_{i,p}$ proportional to the conditional probability that *a worker* is employed in industry i, *given* she works for a firm that exports the product p. $_{c,p}$:] Weighted average of the concentration of employment in our city c across all indutries $i \in \{1,2,\ldots\}$, with weights $w_{i,p}$ proportional to the conditional probability that industry i is present in *a city*, *given* that the city exports the product p. $_{c,p}$:] Weighted average of the intensities (relative to the world) of exports per capita in our city c across all products $p' \in \{1,2,\ldots\}$, with weights $w_{p',p}$ proportional to the similarity between products p' and the product p in terms of how they co-occur with all industries. $_{c,p}$:] Weighted average of the similarities between the product p and all other products $p' \in \{1,2,\ldots\}$ (in terms of how they co-occur with all industries), with weights $w_{p',c}$ proportional to the intensities (relative to the world) of exports per capita in our city c across all products p'.

It is important to notice that $D^{(1)}_{c,p}$ and $D^{(2)}_{c,p}$ measure the relatedness of product p with the industries present in city c, while $D^{(3)}_{c,p}$ and $D^{(4)}_{c,p}$ measure the relatedness of product p with the other products present in city c. If our picture of products being the result of combining ingredients is correct, this difference between densities 1 and 2 versus 3 and 4 may matter, since ingredients may be scarce. Hence, a city may have the right ingredients (i.e., the right industries) to produce product p, but those ingredients may not be available because they may already be in use for other products p' which use the same ingredients as p. That is why we introduce all these different density measures.

Growth of products

The idea is to test whether these densities have any explanatory power in predicting the change in time of variables of interest in a city c for a product p, and in what direction is the effect. The three main dependent variables of interest that we will analyze are the *modRCA*'s, the number of *employees*, and the *number of firms*. In a given regression, then, we will regress the change from a year t to $t + \Delta t$, against the current level of the variable of interest at t, the densities, and some fixed effects that we may or may not want to control for.

In some regressions we are going to be including all the densities in some of the specifications. Thus, we want to anticipate problems of multicollinearity. Below in Table 4 we report the pair-wise correlations between the density variables.

[INSERT TABLE] [!htbp]

Table 4: Pairwise correlations between density variables.

ccccc

As expected, all densities are positively correlated, yet they are not perfect substitutes. Since the highest correlation is between $D^{(1)}$ and $D^{(2)}$, it is important to keep in mind this when interpreting the results.

To run the regressions, we decided to explore (almost) all possible specifications to reduce the risk of "p-hacking", or the so-called "garden of forking-paths" (see `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`). The specifications are all the possible combinations defined by the following pieces:

$$\Delta Y_{t,t+\Delta t} = \beta_0 + \beta_1 Y_t + \mathbf{D_t}\beta + FE, \quad \text{for each city } c \text{ and product } p.$$

1. Y**:** Dependent variable as one of the following **three** options: {modRCA, log(employment), log(number of firms)}. Notice that since each observation in the regression is for a city-product pair, the dependent variables of employment and number of firms are only with regards to exports. In other words, it is the employment and the firms engaged in exporting product p in city c.

2. Δt**:** "Change" of dependent variable defined over a period of time from one of the following **five** options: $\{1,2,\ldots,5\}$ years.

3. D**:** Independent variables as one of the following **five** options: {all four densities, $D^{(1)}_{c,p}$, $D^{(2)}_{c,p}$, $D^{(3)}_{c,p}$, $D^{(4)}_{c,p}$ }.

4. FE**:** Fixed effects as one of the following **four** options: {no F.E., city F.E., product F.E., city F.E. and product F.E. }.

These yield a total of $3 \times 5 \times 5 \times 4 = 300$ different regressions to be run. From the first two options, there are 15 different dependent variables: three types of dependent variables each with five different time windows. Which means that for a single dependent variable, there are 20 different regressions. In all 300 regressions, the specifications that have the least amount of observations (i.e., the smallest sample size) is when we consider 5 year time windows of change in the dependent variable. In those cases, a regression will have approximately 7,000 observations (which comes from the combinations of 62 cities times 617 products and 2 sets of 5 year windows from 2008 and 2014, divided by 10 because only 10% of city-product combinations exist in the data). When all densities and all fixed effects are included, there will be $1+1+4+(62-1)+(617-1)=684$ (the intercept, the reversion to the mean term, the four densities, the cities plus products FEs, respectively) coefficients to estimate. Thus, we will have reasonable statistical power to estimate these regressions.

[INSERT FIGURE]

[h!]

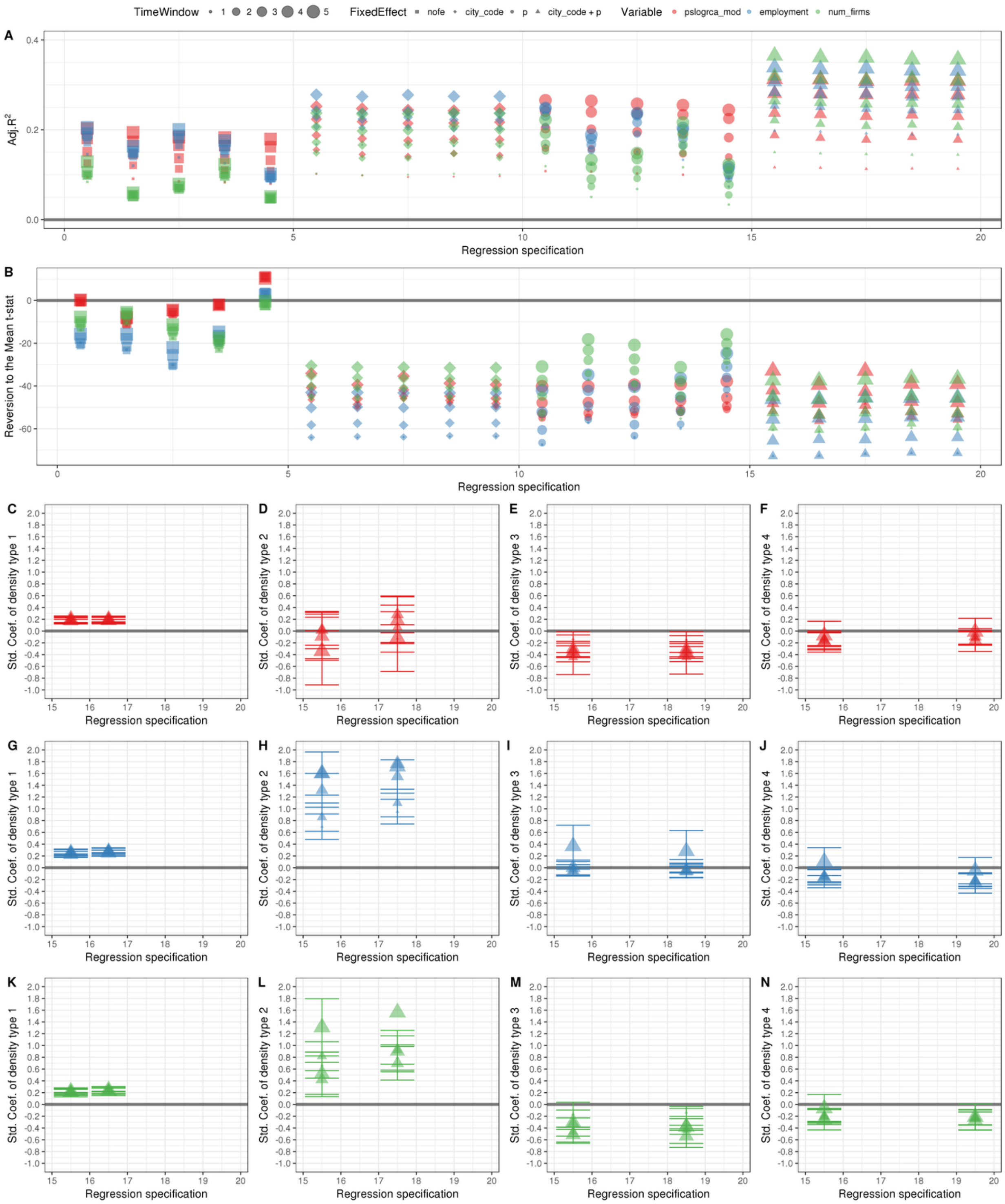
TimeWindow • 1 ● 2 ● 3 ● 4 ● 5
FixedEffect ■ nofe • city_code • p ▲ city_code + p
Variable ● pslogrca_mod ● employment ● num_firms
A
Adj.R^2
0.0
0.2
0.4
Regression specification
0
5
10
15
20
B
Reversion to the Mean t-stat
0
-20
-40
-60
Regression specification
C
D
E
F
G
H
I
J
K
L
M
N
Std. Coef. of density type 1
Std. Coef. of density type 2
Std. Coef. of density type 3
Std. Coef. of density type 4
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
15
16
17
18
19
20
Regression specification

Figure 6: Visualizations of features of the results of 300 regressions (each dot refers to one of the regressions). **Panel A**: Adjusted R^2 of the density regressions. **Panel B**: t-statistic of the term for the reversion to the mean. **Panels C-F**: Estimated coefficients (standardized), with 95% confidence bars, for the four densities when the dependent variable is change in modRCA. **Panels G-J**: Estimated coefficients (standardized), with 95% confidence bars, for the four densities when the dependent variable is change in employment. **Panels K-N**: Estimated coefficients (standardized), with 95% confidence bars, for the four densities when the dependent variable is change in number of firms. In all panels, each value on the x-axis is one of 20 different regression specifications, and for each of the values, there are 15 dots (vertically located since they correspond to the same x value), one dot for each unique dependent variable, which consists of a combination of different time windows (shown as five different sizes), and three three types of dependent variables, modRCA (red), employment (blue), and number of firms (green). **Panels C-N** have separated those 15 values into the different types of dependent variables and that is why colors have been sorted.

Figure 6 shows a total of fourteen plots. The figure synthetizes the most important results from our density regressions. Hence, it is worth going over each piece of the figure in detail.

The x-axis for all plots, **A-F**, is the list of all the 20 regression specifications. For each value in the x-axis we observe several dots scattered vertically, which represent regressions for the 15 different dependent variables. These 15 dots per x-value are given by the three dependent variables and the five different time-windows. Three colors make the distinction between the different types of dependent variables, while the size is given by the "Time Window" used for a specific dependent variable.

The 20 specifications in the x-axis are visually divided by the shapes of the markers in four sets: squares (x values between 1-5), diamonds (x values between 6-10), circles (x values between 11-15), and triangles (x values between 16-20), which correspond to no fixed effects, city fixed effects, product fixed effects, and both city and product fixed effects. Within a given set (i.e., for a specific shape) there are five specifications which, in order, are: all densities included, only density 1 included, only density 2 included, only density 3 included, and only density 4 included.

The first, plot **A**, shows all the adjusted R^2 for all the 300 regressions, each for one of the three dependent variables. We observe that all fixed effects increase the R^2. City fixed effects give the highest performance when predicting the change in employment (blue diamonds in x values between 6-10), product fixed effects give the highest performance when predicting change in modRCA (red circles in x values between 10-15), while including both city and product fixed effects give the highest performance when predicting the change in the number of firms. We also observe, by noticing the size of the dots, that the highest R^2's are reached for time windows of 5 years.

One should notice from the Figure 6, panel **A**, that even when no fixed effects are included (square shapes) the R^2 remain high for these type of regressions (surprisingly, given that we are trying to predict time-changes in very disaggregated variables regarding products within cities).

In panel **B** we show the t-statistic of the term for the reversion to the mean. This plot is meant to confirm that this term should be negative. Overall, it is indeed negative, except for the specification with no fixed effects and all densities included (x value of 4.5). We observe that the significance of this term slightly increases as we include more fixed effects. Its effect is particularly strong when the regressions are for employment (blue-colored dots).

From the previous observations, based on panels **A** and **B**, we conclude that the most robust specifications are when we include both city and product fixed effects. This means that regressions are best at explaining the future success of each specific product in each specific city *relative* to other cities (exporting the same product) and to other products (exported by the city). Hence, in panels **C-N**, we only show the specifications that include both fixed effects, which is why the values on the x-axis only cover the range between 15 and 20. In those twelve plots, we show the estimated coefficients (standardized) of the four density variables for the three dependent variables. Rows separate the dependent variables into modRCA (red, **C-F**), employment (blue, **G-J**), and number of firms (green, **K-N**), while columns show the estimated coefficients for $D^{(1)}$ (panels **C**, **G**, **K**), $D^{(2)}$ (panels **D**, **H**, **L**), $D^{(3)}$ (panels **E**, **I**, **M**), and $D^{(4)}$ (panels **F**, **J**, **N**). In each individual panel there are only 10 dots: 5 for the specification in which all densities are included, and 5 for the specification in which the density appears alone (and the "5" refers to five different time windows). The size of the dots still corresponds to the time-window used in the regression. The values of the coefficients can be compared across regressions and specifications because they have been estimated on standardized variables.

We start the analysis of the densities by first commenting on the fact that the density type 4, $D^{(4)}$ seems not to be statistically significant (except slightly with a negative predictive effect on the number of firms in a specification where the rest of densities are not included). The other densities, however, have more interesting effects.

From the panels **C**, **G**, and **K**, we can see the values of the estimated coefficients of the density type 1, $D^{(1)}$, is very stable across both specifications and for all types of dependent variables, even for the different time windows. Density type 2, $D^{(2)}$ (panels **D**, **H**, and **L**), seems to be predictive of changes in employment and changes in number of firms (with a strong effect), but not of changes in modRCA's. Both densities $D^{(1)}$ and $D^{(2)}$ show the expected positive sign, and are stable to the inclusion of the other densities, despite the correlations shown in Table 4. The meaning of these results is that the availability of related industrial resources (either workers, or the mere presence of an industry in the city) increases the chances of increasing the exports in a product.

Density type 3, $D^{(3)}$ (panels **E**, **I**, and **M**), however, seems to be negatively related to changes in modRCA and changes in the number of firms across the different specifications, and is not statistically significant when estimating changes in employment. The unexpected negative sign may signify that products compete with each other to be exported in a city. Given that the interpretation of this density is of a weighted average of the modRCAs of a city across *products*, it suggests that a decrease in the possibility of exporting a given product is associated with having presence in other very similar products, where similarity is given by a product's industrial requirements. In other words, given the effects of $D^{(1)}$ and $D^{(2)}$ versus $D^{(3)}$, the story that emerges is that an export product is most likely to grow in a city if (i) there are the relevant industrial resources, *but* (ii) there are no other products already being exported in the city that use those industrial resources. The first effect from $D^{(1)}$ and $D^{(2)}$ is one that drives diversification, while $D^{(3)}$ drives specialization. Which of both effects wins? We know, from looking at the real world, that the first effect must win, since larger cities are more diversified, not more specialized. Currently we do not have a time frame long enough to see whether this effect becomes negligible with longer time-windows.

We end this section by reporting the results of specific regressions that have the highest statistical significance, but also with the clearest economic significance based on the above analysis. The regressions include all city and fixed effects, and predict the changes in modRCA, employment and number of firms, over a 5 year time-window. We include only $D^{(2)}$ and $D^{(3)}$ as our densities of interest. Table **??** corroborate our previous findings indeed that predicting changes in modRCA is harder than changes in employment, and changes in employment are harder to predict that changes in number of firms.

We also see that, at least for employment and number of firms, $D^{(2)}$ is positive and stable, while $D^{(3)}$ is positive for employment change but negative for changes in number of firms, but in all cases it is relatively stable. All regressions have R^2 above 0.3, but this seems to be coming mainly from the reversion to the mean term, and the city and product fixed effects. We find, however, something that Figure 6 did not reveal: when both densities are included in order to predict changes in employment and number of firms, $D^{(3)}$ becomes weakly significant. Hence, it suggests a resolution to the issue between the diversification and specialization effects. We can conclude from both the size of the effects and their statistical significance that products compete over the industrial resources in the city, but the net effect is that diversification processes are stronger. A corrollary of this, is the finding that industries have a certain rival aspect with each other. Rival goods, when used for an activity, cannot be used for something else. This mechanism seems to be what $D^{(3)}$ is picking. Further studies could in principle disentangle in more detail which industries act as rival and which as non-rival (this division has typically been conceptualized as physical capital being rival and human capital being non-rival).

[INSERT TABLE] [!htbp]

Table 5: **modRCA regression** table showing the definitive specification of our densities. All variables have been standardized before the regression, so the estimates are for standardized coefficients. The density $D^{(2)}$ based on the relatedness with industries shows a positive effect on the change in modRCA, while the density $D^{(3)}$ based on the relatedness with existing products in the city shows a negative effect. Standard errors shown in parenthesis.

lcccc

[INSERT TABLE] [!htbp]

Table 6: **Employment regression** table showing the definitive specification of our densities. All variables have been standardized before the regression, so the estimates are for standardized coefficients. The density $D^{(2)}$ based on the relatedness with industries shows a positive effect on the change in employment, while the density $D^{(3)}$ based on the relatedness with existing products in the city shows a negative effect. Standard errors shown in parenthesis.

lcccc

[INSERT TABLE] [!htbp]

Table 7: **Number of firms regression** table showing the definitive specification of our densities. All variables have been standardized before the regression, so the estimates are for standardized coefficients. The density $D^{(2)}$ based on the relatedness with industries shows a positive effect on the change in number of firms, while the density $D^{(3)}$ based on the relatedness with existing products in the city shows a negative effect. Standard errors shown in parenthesis.

lcccc

Appearance of products

The previous empirical exercise investigated the time change of three continuous variables, modRCA, employment and number of firms, in a city c and a product p. We concluded that two of our four densities were significantly associated with those future changes. However, these results were limited only to those cases in which the three variables had already a positive value for the product p. In other words, our results only applied to the cases when there was already *something*. But what if there was *nothing*, instead?

Our previous results do not tell us anything about the situations in which there are no firms (hence no employees and no exported value) related to a product p in city c at time t. The question in this subsection is thus: Are our four densities predictive of *product appearances* (i.e., from “nothing” to “something”)?[6]

Given that firms are the units of production, and given that our previous results cover the growth of exports even if the production is very small, we will concentrate on the following strict defition of appearance (again, for a given city c and product p):

Absolute absence at time t $\longrightarrow$ *At least a firm with at least 1 effective employee at time $t+\Delta t$*

It is important to note that this definition of appearance of a product in a city does not necessarily imply that only *new firms* are responsible for new products in the city; it may be that an appearance is due to an already *existing* firm starting to export a *new* product that no firm before exported until that moment.

We perform conventional logistic regressions (i.e., we fit *logit* models), as

$$\mathrm{A}_{t,t+\Delta t} = \mathrm{logit}(\beta_0 + \mathbf{D_t}\beta + FE).$$

We also introduce a way of studying models which we will use in future sections which is different from the conventional way of looking at a regression models. For each time-window in which we are trying to model appearances we split our data in two: a *training* set and a *test* set. We fit the model and we estimate parameters on the former and we evaluate its predictive power on the latter. We choose our test set as the last observations in time. For example, if we are predicting appearances over 5 year periods, (i) we will take the information on 2008 to compute our densities, (ii) fit a logit model based on how well the densities predict appearances of products from 2008 to 2013, and (iii) based on the fitted model, we will use the information in 2009 to make *out-of-sample predictions* of appearances in 2014.

Our logistic regressions return a predicted probability (a number between 0 and 1) which serves as an indication of whether the product will appear or not. To make this decision, however, one must choose a threshold above which we will be confident of saying that the product will, in fact, appear. But how to choose a threshold if we want to *minimize false predictions*? If we choose a high threshold, we will be predicting only very few appearances, and thus we will minimize the risk of predicting an appearance that won’t happen. Hence, a high threshold will lower our rate of False Positives. Conversely, if we choose a small threshold, we will be predicting lots of appearances, and we will minimize the risk of predicting that something won’t appear when in fact it does. Hence, a low threshold will lower our rate of False Negatives. Clearly, there is a trade-off, and one must sacrifice one or the other, depending on our goal.

[6] These two cases are often referred to as predicting the *intensive margin* or the *extensive margin*, because the former predicts changes in the intensity of an already existing variable, whereas the latter predicts appearances of new elements.

Given this arbitrary choice for picking the threshold for a predicted probability of appearance, the convention is to use the notion of the ROC curve (from “Receiver Operating Characteristic”). This is a curve, given a fitted model, that shows the trade-offs that come from changing that threshold. And then, the convention is to compute the *area under the (ROC) curve*, or AUC. The AUC can be interpreted as an average performance of the fitted model across all thresholds. If AUC=0.5, the model is no better than random guessing. If AUC=1.0, the model is a perfect predictor. Hence, the best models have AUC that are close to 1.0, althought typically AUCs above 0.8 are what characterize good models for prediction.

We carry out 25 different regressions. The reason we do not have 300 as before is twofold. First, we have limited the options to only regressions with both city and product fixed effects, and second, there are not different *types* of dependent variables anymore since now we only have a single binary dependent variable, $A_{t,t+\Delta t}$, that represents appearances. Our only options are the five different time windows for Δt and which densities we include, for which we have the same five options of including all or each of the four separately. Thus, $5 \times 5 = 25$ different regressions. An important caveat is that we will not be using true fixed effects, but rather characteristic variables of the cities and the products instead. The reason is computational. The logic behind fixed effects is not strictly applicable to logistic regression since there is no sense of “additiveness” of the covariates. So, while one can fit a model with many dummies, the results are very unstable, and are very computationally demanding (very often the models do not converge). We solve this problem simply by adding the working age of the city and the ubiquity of the product.

We report our results in Tables **??**, where we show the Akaike Information Criterion (AIC) as quantifying the relative performance of the models, in a way that takes into account the complexity of the model (i.e., the number of parameters). Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are computed from the log-likelihood of each model, and in most cases give the same ranking of models, as in this case, so we do not show BIC.

AIC is like the adjusted R^2 used in linear regressions in that it penalizes those models that have too many variables. However, in contrast with the R^2, the best model is the one that has the *smallest* AIC, and this happens for those models that predict well with few variables, or despite having many variables. The AIC is naturally smaller for small sample sizes (the reason being that the log-likelihood is in turn smaller the larger the sample size), and thus one can only compare AIC between models that use the same number of observations. Accordingly, we have ranked the regressions first by the number of observations, and second by their AIC. AIC, however, must be understood differently from AUC. We will compute the AIC based on the training set, and the AUC on the test set.

[INSERT TABLE] [!htbp]

Table 8: Results from logistic regressions done over a training set consisting of all observations except the last appearance (e.g., if the model is predicting appearances over 5 years, then it is only trained over the change from 2008 to 2013 and the change from 2009 to 2014 is left out). Small Akaike Information Criterion (AIC) values mean the model performed well on the training set. High Area Under the Curve (AUC) values mean the fitted models was highly predictive of appearances in the test set (i.e., out of sample predictions). All the regressions include city working age population and product ubiquity.

c—ccc—cc—cccc

[INSERT TABLE] [!htbp]

Table 9: Same results as Table 8 but showing the z-statistics of the coefficients for the densities. All are positive and large (i.e., statistically significant).

c—ccc—cc—cccc

Non-rivalry stronger than rivalry?

The first conclusion we can draw from Tables **??** is that *all* densities in *all* specifications are statistically significant, with only few less significant values for $D^{(1)}$ and $D^{(4)}$. The second conclusion is that the model always performs best when all densities are included, although only including $D^{(2)}$ performs almost equally well. And the third conclusion is that all densities are *positively* predictive of appearances, in contrast with what we found for growth where $D^{(3)}$ had a negative effect (although we stress that adding working age population and product ubiquities may not substitute city and product fixed effects, so the effects of very diverse cities or very ubiquitous products may not be totally accounted for).

When all densities were included, the AUC, i.e., the predictive power for out-of-sample data, was always above 0.825. These are highly predictive models. We note that the highest was for regressions # 11 and # 21, according to the table, althought the digits not shown reveal that the best was really predictions for 1 year periods.

We show in Figure **??** the corresponding ROC curve for the 1-year-period appearance model with all densities included. The x-axis is the "specificity", which is another word for "true negative rate", and the y-axis is the sensitivity, or the "true positive rate". One can see that the model does well because it maximizes both the true predictions. For this model, if we choose the threshold that achieves the maximum sum of specificity and sensitivity, we get a threshold of 0.02. This means that when our model estimates a probability of appearance above 0.02, we will say that the product will appear, and if it's below that value, we will say it will not appear.

[INSERT FIGURE]

According to this threshold, we can construct the specific matrix that counts when the model predicted correctly and incorrectly the appearances and the lack of appearances. This is called the "confusion" matrix. We show that in Table 11.

[INSERT TABLE] [!htbp]

Table 10: Confusion Matrix for the model with the highest $AUC = 0.83$, and for the specific threshold probability 0.02, which maximized specificity and sensitivity. TN = "true negative"; FN = "false negative"; FP = "false positive"; TP = "true positive".

ccc

The confusion matrix shows that there were 479 product appearances from 2013 to 2014 across all 62 cities. Only 88 of those (18%) we incorrectly labeled as "not appearing". Hence, we correctly predicted 81.6% of actual appearances (our sensitivity). On the other hand, we predicted a total of $9,554$ appearances, but only 391 materialized as correct predictions (4% of our predicted appearances). Of all the cases in which nothing appeared (37403 cases) we correclty predicted 75.5% of those (our specificity). Overall, our accuracy (how many "trues", regarding both appearances and lack of them, over the total possible observations) was of 76%.

In conclusion, similar mechanics are behind the growth and appearance of export products in cities. The strongest effect typically comes from density 2, which captures the role of industries, while the next effect comes from density 3, which captures the presence of other products which use those industries.

[h!]

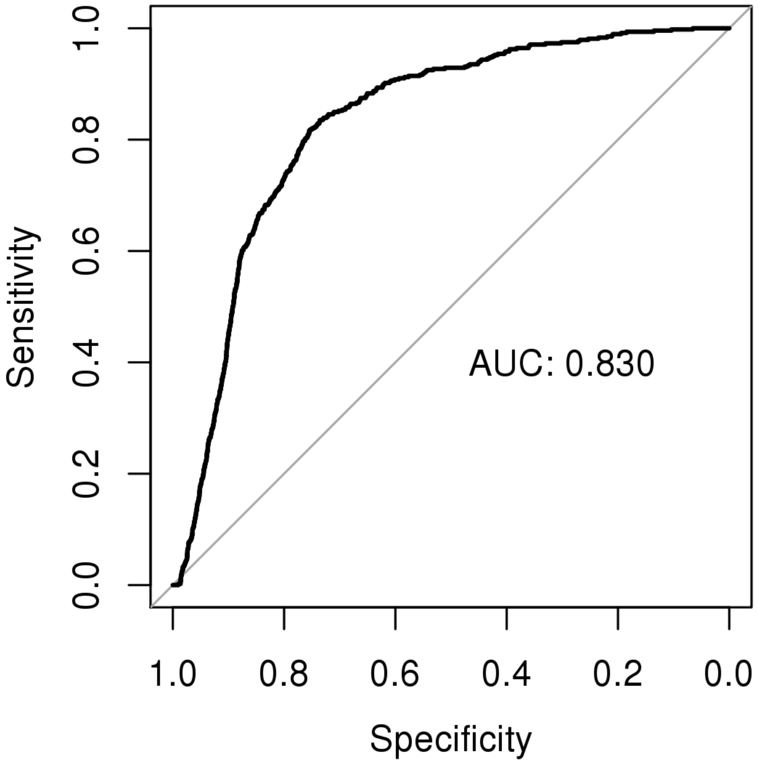

Figure 7: ROC curve over test set for predicting product appearances in cities from 2013 to 2014, having fitted a logistic model for all previous 1 year transition periods.

According to this threshold, we can construct the specific matrix that counts when the model predicted correctly and incorrectly the appearances and the lack of appearances. This is called the "confusion" matrix. We show that in Table 11.

[INSERT TABLE] [!htbp]

Table 11: Confusion Matrix for the model with the highest $AUC = 0.83$, and for the specific threshold probability 0.02, which maximized specificity and sensitivity. TN = "true negative"; FN = "false negative"; FP = "false positive"; TP = "true positive".

ccc

The confusion matrix shows that there were 479 product appearances from 2013 to 2014 across all 62 cities. Only 88 of those (18%) we incorrectly labeled as "not appearing". Hence, we correctly predicted 81.6% of actual appearances (our sensitivity). On the other hand, we predicted a total of $9,554$ appearances, but only 391 materialized as correct predictions (4% of our predicted appearances). Of all the cases in which nothing appeared (37,403 cases) we correclty predicted

75.5% of those (our specificity). Overall, our accuracy (how many "trues", regarding both appearances and lack of them, over the total possible observations) was of 76%.

In conclusion, similar mechanics are behind the growth and appearance of export products in cities. The strongest effect typically comes from density 2, which captures the role of industries, while the next effect comes from density 3, which captures the presence of other products which use those industries.

Summary

We list below a synthesis of what we have learned up until this point about the linkages between industries and exports, specifically based on Section b and this section:

- There are **large differences between firms** in Colombia in terms of their employment size, their total exports, and the number of products they export (Table 1). This matters because aggregates at the level of cities may show fluctuations that are really the reflection of changes in a few individual large firms.
- The distributions across firms of employment sizes, exports, and number of products are lognormal (Figures **??**). This suggests there are factors that induce growth **multiplicatively**.
- The data suggests that **product diversification** is the main driver of firm growth (Table **??** and Figure **??**).
- Cities, on average, export competitively products in which they have **high concentration of employment**. More specifically, concentrations of employment in export products (relative to the national expectation) correlate, on average, strongly with the international competitiveness in export values per capita, i.e., relative to the international reference (Figure **??**). However, there is still large variation around the relationship, so there are cities with low concentrations of employment in a product which nevertheless are internationally competitive, and viceversa, there are cities with larger-than-expected concentrations of employment in a product whose exports are not internationally competitive.
- There is an exponential relationship between the number of products a city exports (the *product diversity*) and the number of industries a city has (the *industry diversity*), observed in Figure **??**, Table **??**, and Equation b. This confirms the model in which **industries act combinatorially** to generate different export products.
- Products and industries, however, do not form well-defined joint clusters, as revealed by the lack of communities in Figure **??**, which prevents the construction of a Product-Industry Joint Space.
- Our *measures of density* which quantify the presence, in a city c, of industries related to a product p are predictive of the growth, over different time windows, of that product (in competitiveness, employment, and number of firms). On the other hand, the presence of other products, which are themselves related to industries in the same way that p is, has a

negative effect on the growth of product *p* (Figure 6). This suggests that industrial employment is a rival good for export production. On the net, however, the **presence of the "right" industries tend to foster the growth of exported products** (Tables **??**).

- Predicting the growth over periods of **5 years** is easier than over periods of 1 year.

- Our measures of density are successfully able to **predict the *appearance* of firms exporting a product *p* in a city**, i.e., from "nothing" to "something" (Table 8), given knowledge of the industries and other products present in the city. The best predictions were found for 1 year time windows with an AUC of 0.83.

We now have a clear picture of the fact that industrial presence *is* a determinant for export diversification. However, results like these are still far from being reliable indications to base public policy decisions on. Instead, our results invite us into further explorations about whether we can actually increase our predictive power by being more agnostic about how exactly industries act together to induce exports (our densities all assume additive linear associations).

[1] C. I. Jones, P. M. Romer, The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital, American Economic Journal: Macroeconomics 2 (2010) 224–245.

[2] P. M. Romer, Increasing Returns and Long-Run Growth, The Journal of Political Economy 94 (5) (1986) 1002–1037.

[3] P. M. Romer, Endogenous Technological Change, Journal of Political Economy 98 (5) (1990) S71–S102.

[4] R. E. Lucas Jr., On the mechanics of economic development, Journal of Monetary Economics 22 (1) (1988) 3–42, doi:10.1016/0304-3932(88)90168-7.

[5] J. E. Rauch, Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities, Journal of Urban Economics 34 (3) (1993) 380–400, ISSN 0094-1190, doi:http://dx.doi.org/10.1006/juec.1993.1042, URL http://www.sciencedirect.com/science/article/pii/S0094119083710429.

[6] E. L. Glaeser, A. Saiz, The Rise of the Skilled City, Brookings-Wharton Papers on Urban Affairs 2004 (1) (2004) 47–105.

[7] L. M. A. Bettencourt, J. Lobo, D. Strumsky, Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size, Research Policy 36 (1) (2007) 107–120, doi:10.1016/j.respol.2006.09.026, URL http://www.sciencedirect.com/science/article/pii/S0048733306001661.

[8] E. L. Glaeser, J. A. Scheinkman, A. Shleifer, Economic growth in a cross-section of cities, Journal of Monetary Economics 36 (1) (1995) 117–143.

[9] Z. J. Acs, Innovation and the Growth of Cities, Edward Elgar Publishing, Cheltenham, UK, 2002.

[10] G. Duranton, Delineating Metropolitan Areas: Measuring Spatial Labour Market Networks Through Commuting Patterns, in: Advances in Japanese Business and Economics, Springer Japan, 107–133, doi:10.1007/978-4-431-55390-8_6, URL https://doi.org/10.1007/2F978-4-431-55390-8_6, 2015.