How Industry-Related Capabilities Affect Export Possibilities

Final report submitted to Bancoldex and National Planning Department as part of Datlas 2.0 Project

Ricardo Hausmann
Principal Investigator
Director of the Center for International Development
Harvard University
Cambridge, Massachusetts

Authors:[1]
Sid Ravinutala
Andres Gomez-Lievano
Eduardo Lora

CENTER FOR INTERNATIONAL DEVELOPMENT

GROWTH LAB

cid.harvard.edu

Center for International Development at Harvard University

[1]We acknowledge many useful comments made by Bancoldex and DNP technical staff at the Seminar held in Bogotá, July 12-13, 2017. We also received useful comments and suggestions from Frank Neffke, Dario Diodato, Ljubica Nedelkoska, Michele Coscia and Matte Hartog.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Question

The central question we will explore in this document is: Can we anticipate the opportunities that Colombian cities have to export specific products based on their existing productive capabilities?

1.2 Context

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. The individual pieces of know-how are difficult to observe, however.

Using an analogy from biology, the pieces of know-how in a region are like the genes in an organism, while the set of products that the region produces is like the phenotype (i.e., how the organism looks like physically). The mapping in biology from genes to phenotype is subtle and complex. Similarly, the mapping in economics from the set of pieces of know-how in a region to what firms in the region produce is also subtle and complex. However, in biology one can make educated guesses about the genes an organism has based on its phenotype, which, in turn allow one to make predictions about features the phenotype might develop in the future. In the same way, we can make educated guesses about the know-how contained in a region from what its firms produce. And having information about current production is informative about what the region *can* produce in the future.

Now, inferring the know-how of regions has typically been done by looking at a single level of the phenotype: that of aggregated exports of 4-digit products that are internationally traded. However, there is an intermediate level between the individual pieces of know-how and the final phenotype of the products exported: the employment that is distributed in a region across industries. In the analogy with biology, knowing about industries would be equivalent to knowing about specific organs in the body of an organism. Organs are still part of the phenotype, technically speaking, but organs are modules that interact with one another and grow somewhat separately from the rest. The collection of all organs and their combined features is what gives rise to the phenotype of the organisms. In the same way, we hypothesize here that it is the combined effects of the individual pieces of know-how and industries which give rise, and facilitate, the production of products that firms can export internationally.

In brief, then, the question of interest has typically been approached at a single level of the production process, while here we want to go deeper into the structure of these production processes. Methodologically,

traditional efforts (as presented in the current versions of the International Atlas of Economic Complexity and the Colombian Datlas) calculate a measure of "density" which quantifies the likelihood that a place exports a product given other products it exports, using only, as a consequence, the "horizontal" relationships between exports. Here we explore the "vertical" relationships, based on the hypothesis that cities first develop a base of employment in domestic sectors, industries and services, which in turn enable the development of products of high sophistication that can compete internationally and are therefore exported to other countries.

1.3 Aim

In the following pages, we report a collection of results, analyses, and advances in which we *assess how industry-related capabilities affect export possibilities*. Our final goal will be to create a single measure that synthesizes all the knowledge and existing information about the productive capabilities of each city, both "horizontal" and "vertical", and that quantifies how competitive a city can be if it aims at exporting a given product it does not yet export.

This document is broken in two main efforts: First, we want to *understand* the "mechanics" of diversification processes. And second, we want to be able to provide recommendations of products that are not produced in cities, but should be. The first effort requires a multitude of analyses, each trying to *describe* the characteristics of firms, of cities, and of the mechanisms that expand the export baskets of places. The second effort requires the development of a statistical model that is accurate when *predicting* the appearances of products in cities. These two efforts, *explaining* and *predicting*, are complementary, but different. Explanations that lack the power of accurately predicting the future are useless in practice; predictions of phenomena for which we lack understanding are dangerous. But together they provide a unified story that can inform policy decisions.

2 Data

The data we will work comes essentially from three sources: PILA for data about sectors with 4-digit ISIC industry codes[2] and employment sizes at the level of firms, ADUANAS for data about exports with 4-digit HS product codes[3] at the level of firms, and DANE for data about working age populations at the level of municipalities. 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).

2.1 Creating the dataset

The raw file from which we create our data for our analysis is the file "R_201542300347182.dta" which is, roughly, a compilation of all imports and exports at the level of trade transactions of firms since 2006 until 2013. We add another file which has 2014 data, named "R_201542302169742_2.txt".

There are two processes that have been separately developed by two CID researchers which take this dataset, clean it, and generate aggregates for the exports of firms across years and across 4-digit products. The researchers are Matte Hartog and Jose Ramon Morales. It may be important to keep this in mind, in case there are other features that both processes do not share. At the end, however, we choose to keep Morales' procedure, which is already in a more appropriate format for our analysis.

2.1.1 Morales' process

The procedure developed by Jose Ramon Morales had a particular goal: to geographically distribute the exports in the Colombian territory. The reason for this is that the ADUANAS dataset only reports the fiscal address of the exporting firm. Hence, we needed a way to infer the regions in which the exports of a firm had been produced. The essence of the procedure was to find the firm in PILA, establish how its employees were distributed geographically across municipalities, and use that information to distribute the total exports per product proportionally to how its employees were distributed geographically.

Morales' procedure use the files "R_201542300347182.dta" and "R_201542302169742_2.txt", as mentioned above. We run the following do-files sequentially: "trade_001_rawtrade.do", "trade_002_PILAfirms.do", "trade_003_matchPILA.do", and "trade_004_noPILA.do". From this, we generate the data file "exp_firm.dta". This dataset has already distributed exports across municipalities. Thus, it consists of the following variables: firm identifier, municipality, 4-digit product code, year, country destination of exports, and the total value of those exports (both in dollars and in pesos).

[2] The ISIC Rev. 3 A.C. codes, which are adapted to Colombian economic activities.
[3] The HS codes 1992 revision.

To this, we *merge* the data "COL_ciey_2008-2014_do600.dta". After this, we manipulate the data such that we get at the end a list of firms, per year, per municipality, with a column "p" that has a product code, a column "ciiu_rev3" with an industry code, a column "X_pr_d" reporting the exports in dollars, and a column "eff_num_emp" with the effective number of employees[4]. Other information include names of the locations and the economic activities.

2.2 The final firm-level dataset

The final dataset contains $8,524,309$ total rows. In them, there are $2,519,960$ unique firms, across 7 years, $1,123$ municipality codes, 62 city codes[5], $1,179$ product codes, and 469 industry codes. We will use this firm-level dataset for two purposes in Section 3. 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.

In this final firm-level dataset, all rows should in principle have industry and municipality codes. There are, however, a minority of observations that have missing values for those codes. Specifically, there are $89,137$ rows with that information missing. When we start the analysis, we will drop those observations. But let us describe what this means in more detail.

The problem is really that some observations (i.e., firms across a few years) have missing values in the municipal code column. It turns out that the vast majority of these observations are about firms that export Petroleum products. Specifically crude, refined and gas. For Datlas we run a separate process to distribute these exports across the municipalities.

However, we do not run that process for two reasons. The first is that these firms have no industrial code (there are 69 different firms in total). Therefore, the exercise of linking industries with products does not apply to these firms. And the second reason is that petroleum products are unique and not apt for the type of analysis that follows (e.g., based on knowledge-based economic activities in which skills are located in the places of production). This problem is evident in the way oil production is ranked in the complexity indicators produced by both the International Atlas and Datlas itself.

In summary, we drop "Coal" (code 2701), "Crude petroleum oils" (code 2709), "Refined petroleum oils" (code 2710), and "Petroleum gases" (code 2711). An additional reason we drop them is because these products are raw materials which represented more than 63% of total Colombian exports in 2014 (see Figure 1) and are very sensitive to fluctuations in price. Hence, movements in price strongly affect the total exports of Colombia and may hide other productive capabilities.

[4]The notion of "effective number of employees" aims to quantify the number of employees in a month on average. It accounts for the fact that not all workers work all year. For example, a firm with two employees that only worked for 6 months will have an effective number of employees equal to $\frac{2\ employees \times 6\ months}{12\ months/per/year} = 1\ effective\ employee$.

[5]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 Duranton G. (2015) "Delineating Metropolitan Areas: Measuring Spatial Labour Market Networks Through Commuting Patterns", in Watanabe T., Uesugi I., Ono A. (eds) The Economics of Interfirm Networks. Advances in Japanese Business and Economics, vol 4. Springer, Tokyo.

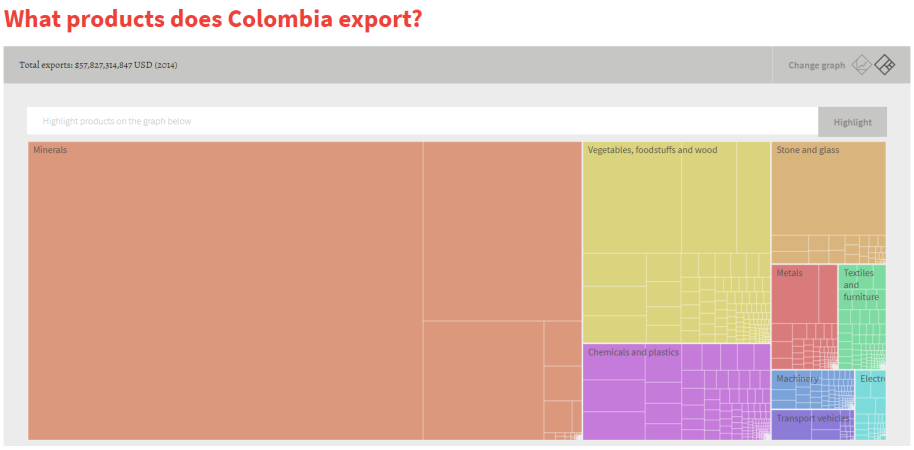

Figure 1: More than 63% of Colombian exports in 2014 consisted of coal, crude and refined petroleum oils, and petroleum gases. In the figure, those are most of the products in the "Minerals" section (brown-orange color).

Firms that lack municipality codes represent a large proportion of the total country exports. There are 56,536 observations in the database in 2014 that had exports (each observation is a combination of firm + municipality + product) in the database. Of these, only 28 have no municipal code. These 28 observations are 25 different firms, while the 56,536 observations are about 7,069 firms. The 56,536 observations represent a total of $53.98 billion (i.e., thousand millions) of dollars exported (Datlas site reports $54.8 billions). But surprisingly, the 28 observations that have no municipality code give a total of $23.21 billion dollars exported. Basically half of total exports. Thus, to be more exact, 0.3% of exporting firms in 2014 exported 43% of all Colombia's total exports. Thus, these are the firms that export the productions we said above that we drop.

After droping these observations we observe that in 2013, according to our dataset, the total number of effective number of employees in the formal sector was 6.88 million workers (Datlas site reports 6.7 million in 2013).

2.3 City-level data

As mentioned, our analysis about the link between industrial employment and export possibilities is focused on Colombian cities. One problem that arises when aggregating the firm-level data to the level of cities for exporting firms is the fact that, on the one hand, we have employment data for firms as a whole in a given location, and on the other hand, we have data about all the several products that a firm exports from the location. What we do not have is how do firms distribute their workforce across the different products they export.

There are at least three possibilities to solve this problem:

1. Approximately 45% of firms (in 2014) exported only one product. The first option is to generate these aggregate employment numbers by city-product only based on that 45% of firms that only export one product.

2. The second option is to try to determine which is the main export product of "multi-product firms", and discard the others. This way we do not lose employees in the total aggregate of the city, but we will not take into account the fact that we are not counting employees that did contribute to some of the products that are secondary to firms.

3. And the third option is simply to distribute the employment of a multi-product firm evenly across all its products (within a specific municipality). That is, if a firm in a city has 10 employees and exports 5 different products, then we say that there are 2 employees per product. In this way all employment aggregates are more evenly distributed across products.

We choose to pursue the 3rd option. It is the one we think discards the least amount of information, and it is at the same time the most agnostic about the use of know-how to produce several products and to export them.

After aggregating the employment, we get to a stage where the dataset consists of combinations of year, city code, industry code, product code, and the effective number of employees associated with each of these combinations. We add to this dataset the working age population from DANE that we have aggregated from municipalities to cities (we say "working age population" to be the population with 15 years or older).

At this point, we could aggregate up one more level in order to get a dataset of employees country-wide by industry code and product code (for each year). In a sense, this would define a matrix of industry-product co-occurrences. The co-occurrence here happens "within individual workers" (as opposed to co-occurrence in a physical location like a city), since we have assigned each worker to an industry and to a product (when the worker works for an exporting firm). To extract the technologically significant connections between industries and exports we would need to control for two issues. The first arises from the fact that some industries employ more people and some products are produced by more people than others, which means that there are industry-product combinations that will have large numbers of people employed merely by chance. To control for this, one would simply divide the actual number by the expected value given random assignments (analogous to a Location Quotient or Revealed Comparative Advantage). The second issue is that many industry codes and product codes pairs exist only because the former is a higher order classification of the latter. In other words, the industry of an exporting firm is, in itself, a broad level classification of the main product the firm produces. Hence, some connections will exist by necessity: firms that export flowers (product code 0603) will themselves report to be part of the flowers sector (industry code 0112), for example. This means that one would need to "subtract" somehow these trivial connections. These issues will be addressed in due time in the following sections.

3 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 1 breaks up these numbers by year, and we report the number of firms per year that do, and do not, export. Some of these exporting firms do not report employees.

	Did the firm export?		
Year	**No**	**Yes**	**Total**
2008	548,263	9,165	557,428
2009	878,607	8,776	887,383
2010	1,156,293	8,035	1,164,328
2011	1,300,385	8,123	1,308,508
2012	1,129,051	8,352	1,137,403
2013	978,245	8,461	986,706
2014	1,030,942	7,076	1,038,018

Source: PILA and ADUANAS.

Table 1: Number of unique firms per year, split by whether they exported or not.

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 2 shows the basic descriptive statistics for the year 2014, for firms that exported.

Table 2: Descriptive statistics of exporting firms in 2014.

Statistic	N	Mean	St. Dev.	Min	Median	Max
Exports (USD)	6,597	$ 7,930,479	$ 217,387,613	$ 0.350	$ 82,990	$ 16,866,765,824
Eff. Employment	6,597	138.7	682.8	0.1	23.9	35,739.6
No. products exported	6,597	4.3	8.5	1	2	167
Average wage (COP)	6,597	$ 20,333,990	$ 20,464,189	$ 400,000	$ 13,880,380	$ 295,634,519

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.

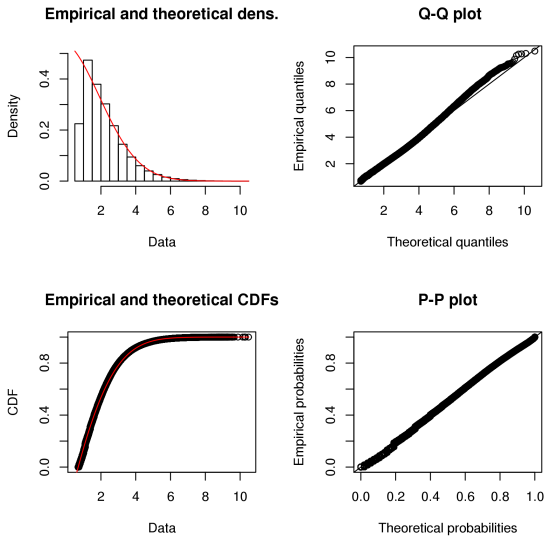

Figure 2: Best fit of a truncated lognormal probability function for the distribution of the number of employees per firm in 2014 based on maximum likelihood.

3.1 Distributions

Figures 2 and 3 plot how the employment size and total exports of firms in 2014 were distributed according to their maximum likelihood fitted distribution function (from a number of candidate distribution functions we tried). Each figure shows four panels: the histogram overlaid with the fitted density function (top-left), the cumulative distribution function (bottom-left), the q-q plot (top-right), and the p-p plot (bottom-right). Both figures have transformed the variables of interest using the natural logarithm. Hence, the normal histograms reveal that the logarithm of the variables look like a normal. But this is precisely because the best fit was a lognormal distribution. For the case of firm size, the distribution is really a *truncated lognormal*, and it is truncated not only because we take only firms that have more than two effective employees, but it seems to be fundamentally truncated. In other words, a multiplicative process that drives the growth of firm sizes would generate a lognormal distribution. However, there would be a natural filtering effect, in that firms of effective sizes below 1 or 2 die easily. The surviving firms would still maintain a lognormal distribution of sizes, but truncated.

Figure 4 plots the same diagnostic graphs for the best fit for the distribution of the number of products exported by firm. The best fit is, again, as for the distribution of employment size, a truncated lognormal distribution. The difference this time is that there are some strong deviations in the right tail. In particular, the q-q plot says that the most diversified firms have less products than one would predict based on the lognormal distribution fitted. Hence, there is a cutoff.

In contrast with the distribution of the number of products exported by firms, which decays rapidly for large firms, the distribution of employment decays more slowly. It is not very apparent in fig. 2 because lognormal distributions are typically confused with Pareto distributions, especially in the right tails. In

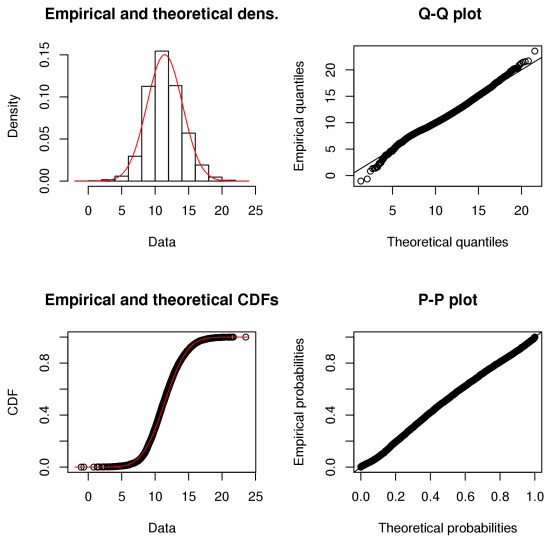

Figure 3: Best fit of a lognormal probability function for the distribution of the total exports per firm in 2014 based on maximum likelihood.

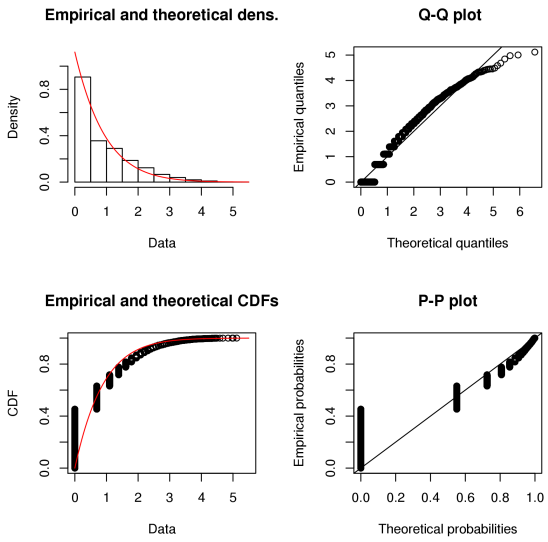

Figure 4: Best fit of a truncated lognormal probability function for the distribution of the number of exported 4-digit products per firm in 2014 based on maximum likelihood.

fact, both distributions can coexist. The variable of the (effective) number of employees is so skewed that instead of plotting the conventional histogram, we plot in the next plot the countercumulative distribution function, taking only firms with more than one effective employee. This is shown in Figure 5. The tail of this distribution seems to be distributed as a Pareto. More specifically, the maximum likelihood fit suggests there is a natural scale of 21 effective employees above which the size of firms is Pareto distributed with an estimated exponent of $\widehat{\alpha} = 2$. Only 4% of firms in 2014 had effective employment sizes larger than 21. We check whether a similar behavior was present in the distribution of total exports but we did not find that

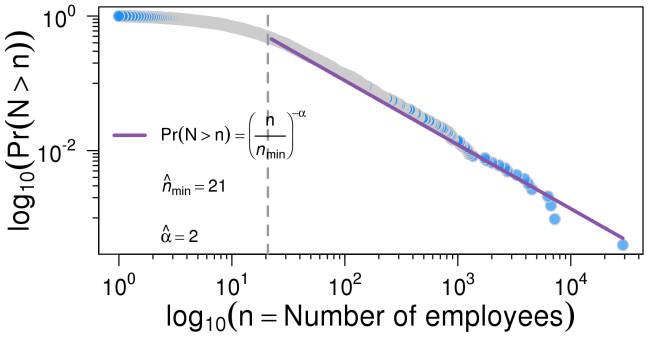

Figure 5: Countercumulative empirical distribution of the number of employees per firm in 2014, showing a Pareto tail.

exports are Pareto distributed, not even far in the right tail. Rather, it is lognormal at all scales.

The implication of these observations about the heavy-tailed distribution of firm sizes is related to the discussion brought up by Gabaix in his paper "The Granular Origins of Aggregate Fluctuations" in Econometrica (2011). In a few words, the argument is that when we observe changes in total employment sizes at the level of a city, for example, we know that these changes are the sums of the changes in employment across firms. But since we know that there are a few firms that are many orders of magnitude larger than most, we know as a consequence that the changes we observe at aggregate levels such as a city are really driven by the changes in a few very large firms. This effect may mislead us in our interpretations of the results once we try to predict changes in employment across cities, or across city-product combinations. We will try to we keep this issue in mind.

3.2 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 6. 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.

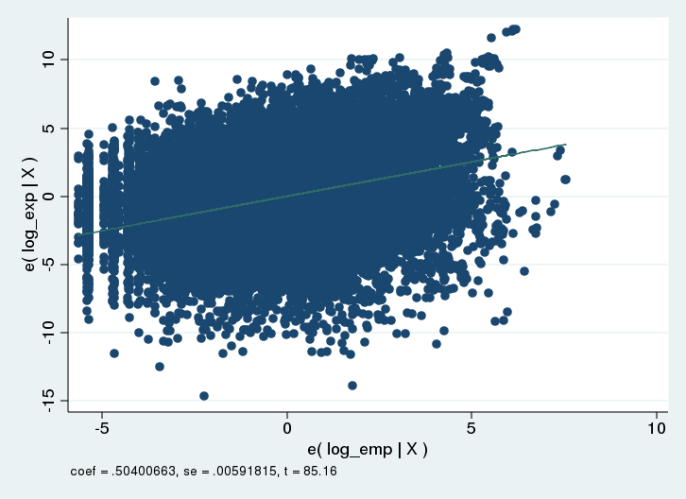

Figure 6: Larger firms export more. This is the scatterplot of the partial correlation of (log) total exports against (log) employment, controlling for year fixed-effects.

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 3 shows the pairwise correlations between these variables for firms in 2014. In addition to the pearson pairwise correlations of Table 3, in Table 4 we show the pairwise *elasticities*.

Table 3: Pearson correlation between pair of variables across firms in 2014.

	log(Exports)	log(Employment)	log(Average Wage)	log(No. products exported)
log(Exports)	1	0.377	0.150	0.410
log(Employment)	0.377	1	0.164	0.283
log(Average Wage)	0.150	0.164	1	0.157
log(No. products exported)	0.410	0.283	0.157	1

In the table, the row acts like the dependent variable and the column the independent variable, controlling for year fixed-effects. From Table 3 we conclude that the pair of variables that correlate the most are number of

Table 4: 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%.

	log(Exports)	log(Employment)	log(Average Wage)	log(No. products exported)
log(Exports)	1	0.504	0.738	1.217
log(Employment)	0.263	1	0.679	0.568
log(Average Wage)	0.041	0.072	1	0.112
log(No. products exported)	0.147	0.132	0.246	1

products produced with total exports. And from Table 4 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 5). In the even columns of the table, we repeat the regression but we include industry fixed-effects. From Table 5 we observe that the effect of firm

Table 5: Results of regressions with year and industry fixed-effects (years explicitly shown).

	OLS regressions							
	Dependent variable: log(Exports)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(Employment)	.50***	.49***					.36***	.29***
	(.01)	(.01)					(.01)	(.01)
log(No. products exported)			1.22***	1.34***			.99***	1.12***
			(.01)	(.01)			(.01)	(.01)
log(Average Wage)					.74***	.98***	.25***	.49***
					(.02)	(.02)	(.02)	(.02)
FE year 2009	−.11**	−.13***	−.10**	−.13***	−.16***	−.20***	−.12***	−.17***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.03)
FE year 2010	−.28***	−.31***	−.22***	−.27***	−.35***	−.43***	−.28***	−.35***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.04)
FE year 2011	−.18***	−.20***	−.15***	−.18***	−.31***	−.37***	−.20***	−.26***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.04)
FE year 2012	−.17***	−.18***	−.09**	−.12***	−.31***	−.37***	−.17***	−.23***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.03)
FE year 2013	−.20***	−.20***	−.10**	−.11***	−.38***	−.43***	−.20***	−.26***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.03)
FE year 2014	−.22***	−.22***	−.06	−.09**	−.34***	−.42***	−.25***	−.31***
	(.04)	(.04)	(.04)	(.04)	(.05)	(.04)	(.04)	(.04)
Constant	10.12***	12.28***	10.55***	13.25***	−.40	−2.38***	5.61***	4.63***
	(.03)	(.32)	(.03)	(.30)	(.31)	(.46)	(.28)	(.40)
Industry FE		YES		YES		YES		YES
Observations	47,553	47,553	47,553	47,553	47,553	47,553	47,553	47,553
Adjusted R^2	.13	.30	.18	.39	.03	.24	.25	.44
Note:								*p<0.1; **p<0.05; ***p<0.01

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 7. 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".[6]

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

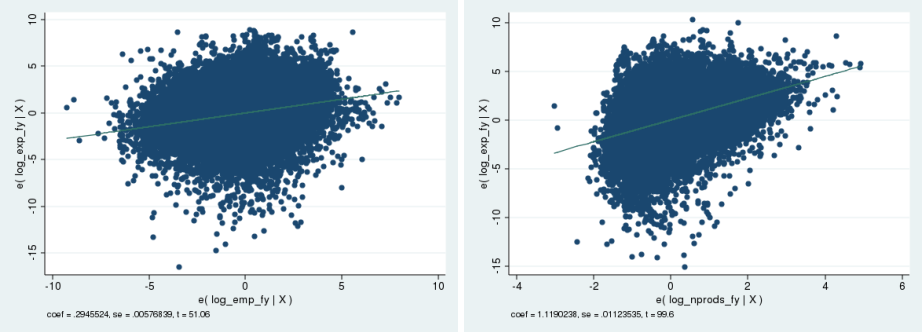

Figure 7: Larger firms export more, and more products. The plots show the partial correlations controlling for all other covariates in the column (8) of Table 5.

3.3 Presences of industries and exports in cities

3.3.1 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.[7] 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.

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

[7] 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.

the statistical empirical distribution of RCAs is lognormal (see fig. 8), 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:

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

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:

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

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 Equation (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 $\text{approxlog}(0,n) < \ln(x_{\min})$, 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 Equation (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.

Figure 8: Comparison of two different ways of calculating RCA-type metrics to quantify the concentration of exports across products in Colombian cities. **Left panel**: top histogram shows RCAs of exports taking as a reference national standards, using a similar formula as equation (4), whereas bottom histogram shows RCAs of exports taking as reference international standards, using formula equation (5). All values in these histograms are for 2014. **Center panel**: scatter plot of the RCAs of exports in cities across all years, where the x-axis uses equation (4) (national competitiveness) while the y-axis uses equation (5) (international competitiveness) with the colors showing the relation across different years. **Right panel**: the same as the center panel, but the colors show the different group categories of the products.

3.3.2 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 Equation (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. Equation (5) will allow us to identify the places that export products competitively with respect to international standards.

We also compute the same exact formula Equation (4) for products, using the formal employment associated to the exports of a specific product in a city using as reference the share of employment in the product nationally with respect to the total national working age population. In what follows, c will always refer to cities. Which way of calculating RCA for products should we use? Are they comparable? Does it matter? Figure 8 compares the histogram of these two measures of *RCA* values for exports for the year 2014, and the relationship between both quantities emphasizing the changes across years (center panel) and the groupings according to 1-digit product codes (right panel). As is very clear from the histograms, being "internationally competitive"[8] is hard. In fact, only 10% of all city-product combinations (for which there

[8]We use quotes to indicate, as mentioned before in the main text, that RCA-type measures do not strictly speaking reveal real "competitiveness". They only reveal relative concentrations of a quantity.

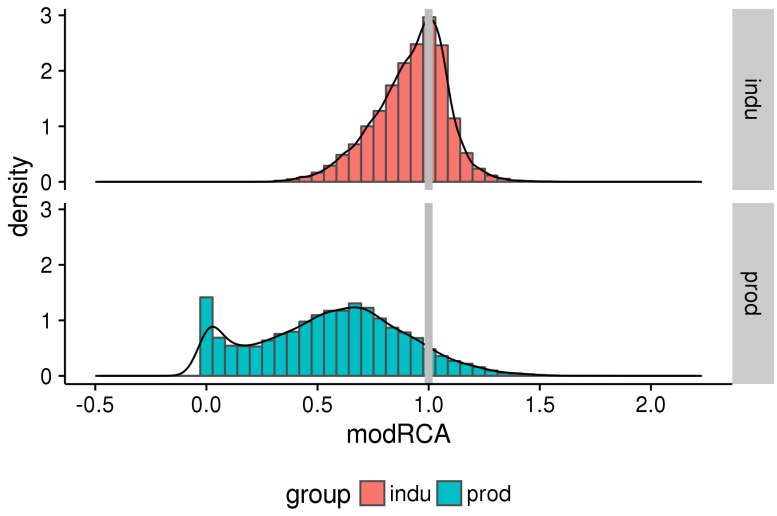

Figure 9: Histogram of *modRCA*'s for all cities and all industries (top plot) and all products except petroleums (bottom plot), which are a transformed version of *RCA*'s but rescaled such that they are now approximately normally distributed. The vertical gray line divides corresponds to $RCA = 1$. Zeros not shown.

are positive exports) reach international competitiveness. However, what is also clear from the scatter plots is that becoming more competitive nationally also makes you more internationally competitive.

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 equation (5) for exports, but we will maintain the formula equation (4) when analyzing industries.

Regarding the distribution of *modRCA*'s that result from transforming equations (4) and (5) using equation (3), we show in Figure 9 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.

3.4 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 co-ordination of productive capabilities, and this process has as a consequence a pattern of diversification, not specialization.

Below in Figure 10 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 fig. 10) is the mean of the *modRCA* (see equation (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 fig. 10 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.

In fig. 10 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.

3.5 Ubiquities per industry and per export

We observe from Figure 10 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 fig. 10. Thus, by computing the so-called "colsum" of a presence matrix we get the vector of ubiquities. This method works well in general, but

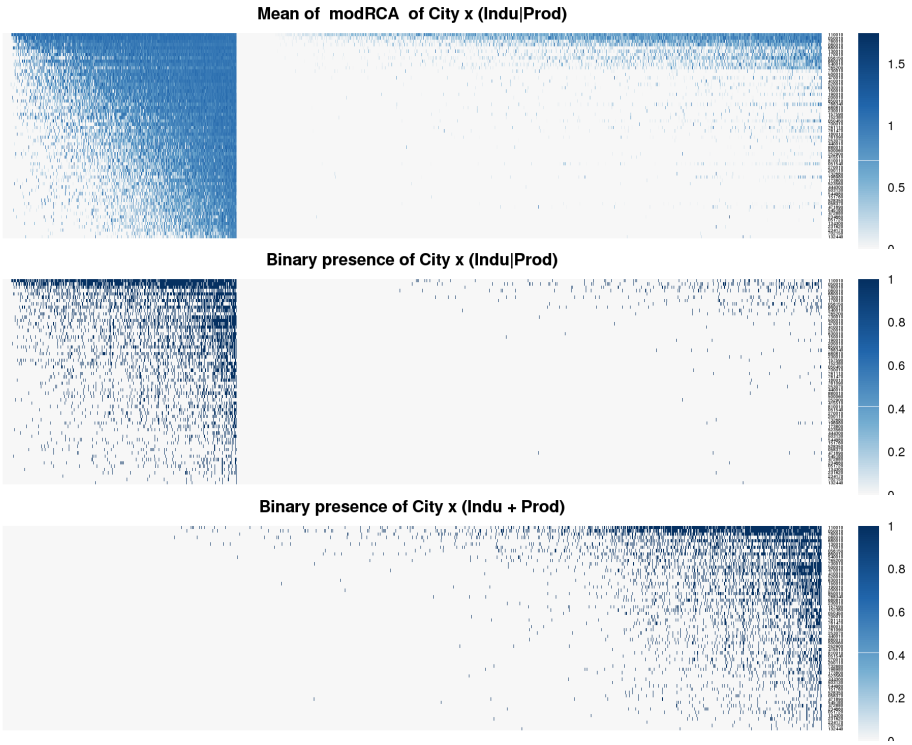

Figure 10: Matrices of average presences across years. All rows have been re-ordered such that the top-most city has the largest number of industry and product presences (together) and bottom-most city the fewest. **Top**: shows the *modRCA*'s (calculated at the city level within Colombia) for industries and products separately, and the columns are ordered so that the most ubiquitous are to the left. **Middle**: The same as the Top panel, but only showing the presences with RCA larger than 1. **Bottom**: the same as the Middle panel, but industries and products are no longer separated, and are ordered simply by the number of cities in which they appear. (Recall that the RCA's for industries have a different interpretation than the RCA's for products. The former is with respect to national employment, while the latter is with respect to worldwide export values per-capita.)

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 fig. 9). 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

$$Ubiquity_i = \sum_c modRCA_{c,i}, \tag{6}$$

$$Ubiquity_p = \sum_c modRCA_{c,p}. \tag{7}$$

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 (6) and (7) are a compromise

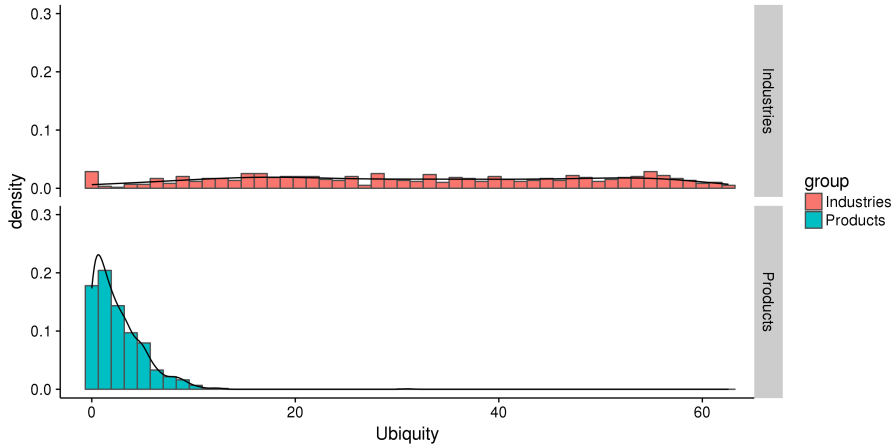

Figure 11: Histogram of ubiquities for industries (top panel) and products (bottom panel).

between these two extremes.

Figure 11 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 fig. 11 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. 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).

3.6 Diversities of cities, with respect to industries and products

We observe from Figure 10 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:

$$InduDiversity_c = \sum_i modRCA_{c,i}, \tag{8}$$

$$ProdDiversity_c = \sum_i modRCA_{c,p}. \tag{9}$$

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

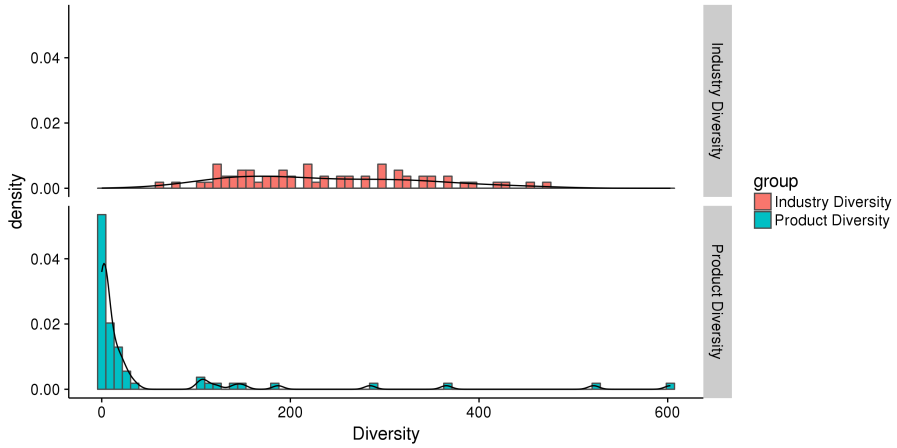

Figure 12: Histogram of industry diversity (top panel) and product diversity (bottom panel).

In Figure 12, 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 fig. 12 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 13, 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 $y = x$.

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

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

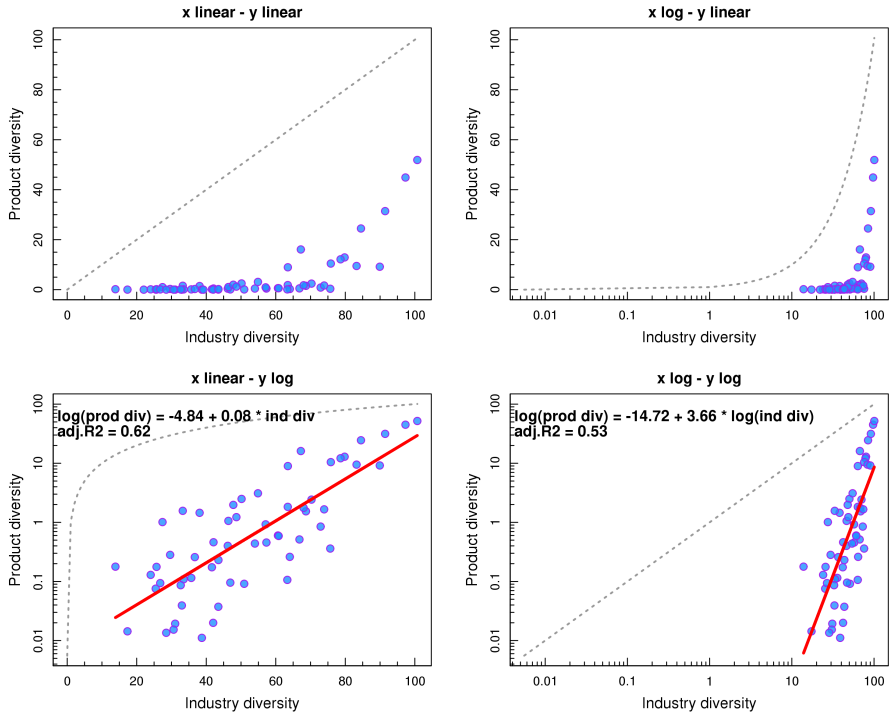

Figure 13: Scatter plot of diversities per city, where each panel has different scales of the axes, to reveal whether there is a linear (top-left), logarithmic (top-right), exponential (bottom-left), or power-law (bottom-right) relationship. The gray dotted line represents the identity line (number of 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.

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 6 shows the results of these fitted regression models.

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

Table 6: Non-linear associations between number of industries and products in cities

	Dependent variable:	
	log(Product Diversity Share)	
	OLS Exponential	OLS Power-Law
	(1)	(2)
Industry Diversity Share	0.082^{***}	
	(0.008)	
log(Industry Diversity Share)		3.658^{***}
		(0.450)
Constant	-4.843^{***}	-14.716^{***}
	(0.480)	(1.756)
Observations	58	58
R^2	0.625	0.541
Adjusted R^2	0.618	0.533
Residual Std. Error (df = 56)	1.363	1.507
F Statistic (df = 1; 56)	93.171^{***}	65.997^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

export products.[9] 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.[10]

Suppose now that a city has an industrial diversity D_I. With that many industries, there are $\binom{D_I}{n}$ possible combinations (i.e., products) of n ingredients (i.e., industries). However, only a fraction q^n will exist. Thus, there will be on average $\binom{D_I}{n} q^n$ products of complexity n in a city with D_I industries. Hence, on average, the product diversity a city will have, D_P is

$$\begin{aligned} D_P &= \sum_{n=0}^{D_I} \binom{D_I}{n} q^n, \\ &= (1+q)^{D_I}. \end{aligned} \tag{10}$$

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

$$\begin{aligned} D_P &= \mathrm{e}^{D_I \ln(1+q)}, \\ &\approx \mathrm{e}^{q\, D_I}. \end{aligned} \tag{11}$$

We see, thus, that product diversity in Equation (11) 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

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

[10] 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.

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

$$\frac{D_P}{M_P} \approx \frac{1}{M_P} \mathrm{e}^{(q\ M_I)\frac{D_I}{M_I}}, \tag{12}$$

or more compactly

$$s_P \approx \frac{100}{M_P} \exp\left\{\left(\frac{q\ M_I}{100}\right) s_I\right\}, \tag{13}$$

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 (fig. 13 and table 6). If we believe this model, the estimates from table 6 tell us that the probability that a product requires an industry is $\hat{q} \approx 0.018$.

4 Mechanics of Urban Export Diversification

4.1 Context

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.

4.2 Aim of this section

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.

4.3 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.

4.3.1 Criteria

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

- Has a ubiquity (equation (6)) larger or equal than the top 95th percentile of industry ubiquities according to their $RCA_{c,i}$ of employment over working age population (equation (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 (equation (6)) 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 Section A 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 (equation (7)) smaller or equal than 2, according to its $modRCA_{c,p}$ (equation (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 Section A 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 equation (7) 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 next subsection.

4.3.2 Effects on employment, number of firms, and export value, of our dropping of industries and products

After trimming the set of all industries and products in our datasets, we are left with 421 industries and 621 products. By dropping observations, some firms are dropped that were associated with some of these industries or products. After this trimming we are left with 543,896 unique firms across all years (recall we originally had a total of 2.5 million firms). In 2014, we are left with approximately 300,000 firms (from one million originally), out of which only 3,000 were involved in exports (originally seven thousand).

The tables below show the effects after dropping both industries *and* products, separately in terms of industries or products.

Table 7 shows what dropping these industries and products entails in terms of employment and number of firms dropped per industry section (i.e., 1-digit industry codes) and year. In the table we see that the largest percentages of total annual employment and number of firms dropped occur in the sections "Financial & Business Services" and "Social Services".

In terms of totals per year, Table 8 shows that we drop approximately half of formal employment (which makes sense, since we are dropping many services), and about 70% of all firms (which also makes

Table 7: Employment and number of firms dropped within industry sections, per year. Each percentage is taken with respect to the national total in the year.

Industry section name	Year	Empl. total	Empl. dropped	% ann. empl. dropped	Firms total	Firms dropped	% ann. firms dropped
Agriculture	2008	180,826	75,415	1.6	9,243	3,818	0.7
Agriculture	2009	192,154	79,508	1.5	10,882	4,234	0.5
Agriculture	2010	198,398	76,163	1.3	12,610	4,883	0.4
Agriculture	2011	191,047	68,976	1.1	13,864	5,156	0.4
Agriculture	2012	205,115	79,397	1.2	14,403	5,223	0.5
Agriculture	2013	216,895	85,715	1.3	14,460	4,731	0.5
Agriculture	2014	245,535	94,471	1.2	17,559	5,609	0.5
Commerce	2008	727,954	348,254	7.2	105,574	52,109	9.4
Commerce	2009	791,689	385,015	7.1	130,648	66,581	7.5
Commerce	2010	859,205	420,279	7.0	153,816	81,484	7.0
Commerce	2011	884,325	433,314	7.1	166,403	87,428	6.7
Commerce	2012	981,433	477,489	7.3	166,888	87,232	7.7
Commerce	2013	1,027,573	493,128	7.2	152,837	76,910	7.8
Commerce	2014	1,180,836	558,582	7.2	165,631	81,719	7.9
Construction	2008	234,033	119,628	2.5	16,219	5,651	1.0
Construction	2009	267,951	142,749	2.6	20,529	6,947	0.8
Construction	2010	305,850	166,664	2.8	23,751	8,014	0.7
Construction	2011	348,223	193,646	3.2	28,804	10,463	0.8
Construction	2012	439,038	247,786	3.8	33,544	13,195	1.2
Construction	2013	516,576	282,960	4.1	42,688	18,188	1.8
Construction	2014	651,381	350,493	4.5	53,722	23,836	2.3
Financial & Business Services	2008	1,509,613	888,413	18.3	230,746	197,894	35.7
Financial & Business Services	2009	1,632,405	968,958	17.9	347,653	310,010	35.1
Financial & Business Services	2010	1,853,564	1,147,346	19.1	526,507	483,713	41.6
Financial & Business Services	2011	1,893,519	1,191,597	19.6	650,684	603,057	46.2
Financial & Business Services	2012	1,954,526	1,198,138	18.3	535,709	486,407	42.9
Financial & Business Services	2013	1,978,906	1,172,098	17.2	445,505	395,436	40.2
Financial & Business Services	2014	2,254,658	1,305,338	16.8	460,278	402,551	38.9
Manufacturing	2008	634,462	265,303	5.5	32,127	10,603	1.9
Manufacturing	2009	643,087	257,810	4.8	36,865	11,891	1.3
Manufacturing	2010	664,187	260,943	4.3	40,923	12,972	1.1
Manufacturing	2011	661,138	256,976	4.2	44,606	13,883	1.1
Manufacturing	2012	739,582	290,118	4.4	45,630	14,196	1.3
Manufacturing	2013	771,181	307,261	4.5	47,205	13,962	1.4
Manufacturing	2014	860,843	337,062	4.3	56,780	14,973	1.4
Mining and Oil	2008	69,541	27,203	0.6	2,399	351	0.1
Mining and Oil	2009	76,078	30,263	0.6	2,874	466	0.1
Mining and Oil	2010	88,228	35,762	0.6	3,370	552	0.05
Mining and Oil	2011	101,011	45,622	0.7	3,777	567	0.04
Mining and Oil	2012	109,678	47,483	0.7	4,159	639	0.1
Mining and Oil	2013	107,202	44,293	0.6	4,738	707	0.1
Mining and Oil	2014	114,278	48,276	0.6	5,247	734	0.1
Social Services	2008	1,146,397	762,602	15.7	138,735	109,313	19.7
Social Services	2009	1,433,715	977,752	18.1	313,569	278,760	31.5
Social Services	2010	1,654,487	1,131,977	18.8	378,100	338,063	29.1
Social Services	2011	1,619,255	1,110,816	18.3	373,403	329,859	25.3
Social Services	2012	1,678,108	1,144,155	17.5	308,885	267,430	23.6
Social Services	2013	1,724,287	1,172,593	17.2	249,641	212,707	21.6
Social Services	2014	1,880,247	1,267,527	16.3	246,090	205,386	19.8
Transport & Communications	2008	307,157	9,928	0.2	18,048	253	0.05
Transport & Communications	2009	329,412	9,236	0.2	20,112	216	0.02
Transport & Communications	2010	345,262	10,290	0.2	21,150	195	0.02
Transport & Communications	2011	348,828	12,792	0.2	22,732	192	0.01
Transport & Communications	2012	402,589	15,934	0.2	23,921	200	0.02
Transport & Communications	2013	438,710	17,034	0.2	25,189	176	0.02
Transport & Communications	2014	513,253	17,414	0.2	28,476	185	0.02
Utilities	2008	33,195	3,706	0.1	984	7	0.001
Utilities	2009	36,230	474	0.01	1,081	2	0.000
Utilities	2010	37,336	478	0.01	1,168	3	0.000
Utilities	2011	36,391	581	0.01	1,236	6	0.000
Utilities	2012	41,888	366	0.01	1,319	5	0.000
Utilities	2013	47,680	605	0.01	1,372	5	0.001
Utilities	2014	51,715	675	0.01	1,544	6	0.001

sense, for the same reasons). In general, our results will be determined by industries that are not services sectors.

Table 8: Effects of dropping industry codes on totals of employment and number of firms, per year.

Year	Empl. total	Empl. dropped	% ann. empl. dropped	Firms total	Firms dropped	% ann. firms dropped
2008	4,843,178	2,500,451	51.6	554,075	379,991	68.6
2009	5,402,721	2,851,766	52.8	884,213	679,099	76.8
2010	6,006,518	3,249,902	54.1	1,161,395	929,871	80.1
2011	6,083,737	3,314,321	54.5	1,305,509	1,050,603	80.5
2012	6,551,957	3,500,865	53.4	1,134,458	874,519	77.1
2013	6,829,010	3,575,687	52.4	983,635	722,814	73.5
2014	7,752,745	3,979,839	51.3	1,035,327	734,991	71.0

The next two tables are only for firms that export. Specifically, Table 9 shows what the dropping of these industries and products entails regarding employment, number of firms, and exported value (in millions of US dollars). As can be seen the major percentages dropped in all three measures happens in the product section of "Vegetables, foodstuffs and wood", followed by "Chemicals and plastics".

In terms of totals per year, Table 10 shows similar results to Table 8. This time, however, we are only looking at the employment that participates in exports. Half of that formal employment is dropped. Betwen 50% and 60% of *exporting firms* are dropped, and about 40% of exported value is dropped.

Finally, in Table 11, we show what these products dropped represent from a classifiction point of view. Products are dropped in all 1-digit categories, although many of the dropped products belong to chemicals, electronics, metals and alloys, and textiles.

As we can be seen, this cleaning process is rather drastic; although a lot of data is discarded, we have aimed for this very conservative approach in order to improve the chances that the remaining information allows us to detect strong relations between industries and exports. Future work could relax some of the criteria used in this study.

4.4 How to construct similarity matrices

Using measures of presence, we can count co-presences of products with industries. Recall that a "presence" is typically defined when an industry (or a product) has RCA larger than 1 in a city. However, recall that we have instead opted to use a transformed value of RCA, the "modRCA", which has the same interpretations as RCA, but which lacks extreme-value behavior. Here we re-define "presence" simple as the value of modRCA, and we count co-occurrences between two economic activities simply as the product of their modRCAs (matrices of these "presences" will be denoted by the letter M). The measures we use here of co-presences are not, strictly speaking, counts. But we will show that these will be "generalized counts", in the same way that the measures of diversity and ubiquity we used in the previous sections based on modRCAs were generalized counts.

If $M^{(cp)}$ is the matrix of "presences" of products in cities, and $M^{(ci)}$ is the matrix of "presences" of industries in cities, then the matrix of co-occurrence of products and industries across cities is computed by

Table 9: Employment, number of firms, and exported values (in millions of dollars) dropped within product classification sections, per year. Each percentage is taken with respect to the national total in the year.

Product section name	Year	Empl. total	Empl. dropped	% ann. empl. dropped	Firms total	Firms dropped	% ann. firms dropped	USD$ total	USD$ dropped	% ann. USD$ dropped
Chemicals and plastics	2008	90,527	54,262	10.0	1,906	1,043	16.0	3,040	1,082	6.7
Chemicals and plastics	2009	86,933	50,478	9.3	1,914	1,055	16.4	2,864	1,072	7.4
Chemicals and plastics	2010	94,331	53,812	10.2	1,889	1,030	17.1	3,017	1,011	6.5
Chemicals and plastics	2011	83,872	46,933	8.9	1,862	1,007	16.4	3,524	1,186	6.5
Chemicals and plastics	2012	98,403	55,271	9.4	1,962	1,056	16.5	3,635	1,333	6.8
Chemicals and plastics	2013	107,641	61,426	9.8	1,941	1,029	15.9	3,908	1,458	8.1
Chemicals and plastics	2014	125,042	75,808	11.4	2,001	1,058	17.4	3,857	1,388	7.9
Electronics	2008	40,658	25,195	4.7	747	442	6.8	487	159	1.0
Electronics	2009	34,942	20,758	3.8	755	455	7.1	458	141	1.0
Electronics	2010	41,103	25,382	4.8	747	448	7.4	321	112	0.7
Electronics	2011	37,210	23,627	4.5	727	455	7.4	373	130	0.7
Electronics	2012	46,769	31,395	5.4	790	483	7.6	430	150	0.8
Electronics	2013	49,129	33,279	5.3	820	487	7.5	451	156	0.9
Electronics	2014	56,093	36,914	5.6	793	495	8.1	441	147	0.8
Machinery	2008	73,219	34,101	6.3	1,765	958	14.7	650	216	1.3
Machinery	2009	92,730	43,710	8.1	1,840	988	15.4	585	217	1.5
Machinery	2010	72,614	33,354	6.3	1,654	906	15.0	455	171	1.1
Machinery	2011	91,275	46,188	8.7	1,746	940	15.3	531	222	1.2
Machinery	2012	100,303	53,741	9.2	1,815	949	14.8	564	253	1.3
Machinery	2013	94,546	51,427	8.2	1,823	950	14.7	625	277	1.5
Machinery	2014	103,998	52,023	7.8	1,868	948	15.6	612	254	1.4
Metals	2008	49,178	28,872	5.3	1,108	572	8.8	1,896	547	3.4
Metals	2009	47,697	28,443	5.3	1,161	626	9.7	1,400	354	2.5
Metals	2010	41,802	18,085	3.4	1,028	558	9.2	1,924	378	2.4
Metals	2011	42,743	23,730	4.5	1,105	598	9.7	1,731	353	1.9
Metals	2012	43,154	22,117	3.8	1,145	636	10.0	1,889	370	1.9
Metals	2013	49,240	26,867	4.3	1,159	611	9.4	1,647	332	1.8
Metals	2014	55,694	28,178	4.3	1,124	614	10.1	1,484	313	1.8
Minerals	2008	8,355	3,603	0.7	193	111	1.7	528	117	0.7
Minerals	2009	7,881	4,179	0.8	211	138	2.1	423	204	1.4
Minerals	2010	8,885	3,256	0.6	176	105	1.7	495	98	0.6
Minerals	2011	10,446	4,643	0.9	160	94	1.5	693	169	0.9
Minerals	2012	14,287	8,354	1.4	180	98	1.5	643	131	0.7
Minerals	2013	10,875	5,927	0.9	165	88	1.4	552	121	0.7
Minerals	2014	8,373	3,533	0.5	176	93	1.5	542	206	1.2
Stone and glass	2008	15,317	4,050	0.7	601	326	5.0	1,403	859	5.3
Stone and glass	2009	15,226	5,489	1.0	600	328	5.1	1,880	1,160	8.1
Stone and glass	2010	16,637	5,338	1.0	536	295	4.9	2,429	1,473	9.5
Stone and glass	2011	13,820	5,675	1.1	536	312	5.1	3,248	1,756	9.6
Stone and glass	2012	14,464	4,725	0.8	531	297	4.6	3,980	2,326	11.9
Stone and glass	2013	17,547	6,132	1.0	525	271	4.2	2,745	1,584	8.8
Stone and glass	2014	16,735	4,969	0.7	452	227	3.7	2,062	1,252	7.1
Textiles and furniture	2008	96,404	58,075	10.7	2,106	1,185	18.1	1,895	1,029	6.4
Textiles and furniture	2009	82,583	48,838	9.0	1,941	1,111	17.3	1,236	679	4.7
Textiles and furniture	2010	85,350	53,878	10.2	1,725	935	15.5	1,115	626	4.0
Textiles and furniture	2011	76,318	42,592	8.0	1,734	961	15.6	1,250	698	3.8
Textiles and furniture	2012	89,462	52,446	9.0	1,783	970	15.2	1,237	727	3.7
Textiles and furniture	2013	96,077	56,093	9.0	1,788	978	15.1	1,133	672	3.7
Textiles and furniture	2014	97,869	59,768	9.0	1,538	872	14.3	1,000	582	3.3
Transport vehicles	2008	10,182	6,893	1.3	303	206	3.2	646	585	3.6
Transport vehicles	2009	9,486	6,844	1.3	309	215	3.3	347	289	2.0
Transport vehicles	2010	8,406	5,819	1.1	263	172	2.9	411	368	2.4
Transport vehicles	2011	11,222	8,343	1.6	256	165	2.7	362	314	1.7
Transport vehicles	2012	14,973	12,055	2.1	292	174	2.7	975	923	4.7
Transport vehicles	2013	13,159	8,879	1.4	280	180	2.8	824	773	4.3
Transport vehicles	2014	18,037	8,813	1.3	289	177	2.9	519	476	2.7
Vegetables, foodstuffs and wood	2008	156,892	76,062	14.1	2,633	1,293	19.8	5,572	1,544	9.6
Vegetables, foodstuffs and wood	2009	164,188	86,846	16.0	2,575	1,333	20.8	5,213	1,406	9.8
Vegetables, foodstuffs and wood	2010	156,817	78,690	15.0	2,390	1,190	19.7	5,319	1,480	9.6
Vegetables, foodstuffs and wood	2011	162,262	85,265	16.1	2,445	1,232	20.1	6,584	1,840	10.1
Vegetables, foodstuffs and wood	2012	163,790	84,221	14.4	2,423	1,214	19.0	6,188	1,924	9.8
Vegetables, foodstuffs and wood	2013	187,271	95,287	15.2	2,440	1,217	18.8	6,215	1,882	10.4
Vegetables, foodstuffs and wood	2014	181,060	98,660	14.9	2,295	1,110	18.2	7,014	1,930	11.0

Table 10: Effects of dropping product codes on totals of employment, number of firms and export value (in millions of dollars), per year.

Year	Empl. total	Empl. dropped	% ann. empl. dropped	Firms total	Firms dropped	% ann. firms dropped	USD$ total	USD$ dropped	% ann. USD$ dropped
2008	540,731	291,115	53.8	6,534	3,730	57.1	16,117	6,140	38.1
2009	541,665	295,585	54.6	6,421	3,761	58.6	14,407	5,523	38.3
2010	525,945	277,615	52.8	6,033	3,463	57.4	15,486	5,717	36.9
2011	529,169	286,996	54.2	6,141	3,545	57.7	18,296	6,668	36.4
2012	585,604	324,327	55.4	6,391	3,628	56.8	19,542	8,138	41.6
2013	625,485	345,316	55.2	6,474	3,623	56.0	18,100	7,256	40.1
2014	662,903	368,666	55.6	6,095	3,403	55.8	17,531	6,549	37.4

Table 11: The trimming of low ubiquity products within 1-digit sections in the product classification. The sections are sorted by the percentage of products dropped.

Product section name (1-digit)	# of 4-digit codes within section	Number of 4-digit codes dropped	% of codes dropped within section
Chemicals and plastics	216	142	65.7
Electronics	47	27	57.4
Metals	141	78	55.3
Machinery	169	93	55.0
Textiles and furniture	166	86	51.8
Vegetables, foodstuffs and wood	267	137	51.3
Stone and glass	64	31	48.4
Minerals	58	16	27.6
Transport vehicles	34	7	20.6

multiplying both matrices,

$$J^{(pi)} = (M^{(cp)})^T \cdot M^{(ci)}, \tag{14}$$

where the $()^T$ means the transpose of the matrix. If all the elements of this matrix are divided by the number of cities, each element is thus an estimate of the probability of observing a given product together with a given industry. As mentioned above, we get a matrix with continuous values, but we will continue using the interpretation as if these were discrete counts.

After having computed the matrix of co-occurrence J we are in a position to compute proximity matrices (we will use the expressions "similarity matrix" and "proximity matrix" interchangeably), from proximity matrices we can calculate the densities, and indices of economic complexity.

Let us recall that an element $[J]_{p,i}$ is the number of cities in which the product p co-occurred with industry i. If we divide by the total number of cities, we thus have an estimate of the joint probability that this product p and this industry i co-occur geographically.

We also need the marginal probabilities of observing a product in a city, as well as the marginal probability of observing an industry. These probabilities are simply their respective vectors of ubiquities divided by the number of cities. Let us call the vector of industry ubiquities by $\mathbf{I}$ and the vector of product ubiquities by $\mathbf{P}$.

Let us denote the joint probability of co-occurrence of products and industries by $\Pr^{(J)}(p,i)$ which we estimate as $[J]_{p,i}/N_c$ where N_c is the number of cities in Colombia ($N_c = 62$); the marginal probability of industries by $\Pr^{(I)}(i)$ which we estimate as $[\mathbf{I}]_i/N_c$; and finally, denote the marginal probability of a product by $\Pr^{(P)}(p)$, estimated by $[\mathbf{P}]_p/N_c$.

An alternative version of the co-occurrence matrix J that we will use below in the empirical exercises is one in which the element $J_{p,i}$ is the effective number of employees that are employed in industry i to export product p. That is, $J_{p,i} = \sum_c E_{c,p,i}$. This matrix is simply the joint frequency of workers that work exporting a specific product from a firm belonging to a specific industry, and so it can be seen as a sort of co-occurrence of products and industries within workforces. We will explain everything in terms of co-occurrence within cities, but everything we say about J will apply to both ways of defining these matrices of co-occurrences (i.e., within cities or within workforces).

4.4.1 Is a joint Product-Industry Space possible? Do products and industries cluster together?

In general, given a matrix of similarities between "entities" (e.g., places, people, genomes, books, disciplines, etc.), one can ask whether entities cluster together in higher level groups. For example, given the co-occurrence of products with industries (i.e., $J_{p,i}$), do products and industries cluster together in well delineated groups?

The practical use of finding well-defined clusters is basically to construct useful visualizations. At present, Datlas Colombia provides network visualizations of the export products on the one hand (the "Product Space"), and of the industrial sectors on the other (the "Industry Space"). The purpose of these visualizations is to aid policy makers and entrepreneurs easily identify technological linkages between nodes (i.e., products or industries), by only looking at the network, and the usefulness of these networks is that nodes do not connect randomly; they connect following some clear patterns that respond to the logic behind the Theory of Economic Complexity.

Based on the present study, finding well defined groups of industries that act as providers of specialized know-how to well-defined groups of products would be very useful, in principle, as it would allow us to visualize a joint Product-Industry Space. However, the fact that industries on their own cluster together, allowing the construction of the Industry Space, and the fact that products cluster together allowing the construction of the Product Space, do not necessarily imply that the joint Product-Industry Space will exist, will make sense and will be useful. This joint space is only useful if groups of industries and products jointly cluster together.

As an example, think of a group of people, where half are women and half are men. Suppose we are interested in the patterns of friendship, and suppose there exists friendships within and between sexes. Women may form well defined groups of friendship among themselves on the one hand, and men may also form well defined groups, on the other hand. However, the cross-sex friendships may, or may not, cluster in well defined groups. For example, if the reasons that drive the clustering in one sex are the same as those that drive the clustering in the other sex (e.g., friendships that emerged from working in the same companies), then we should expect to observe that some specific women *also* cluster with some specific men (e.g., the clusters in the example would represent simply the companies). However, it is also likely that the reasons of the clusterings differ between the sexes. The way to reveal whether joint clusters of men and women exist, would be (for example) to pick women, look at the vectors of women's friendships with other men, and check whether these vectors correlate. In the same way, industries may connect with products, but they may not do so in clusters. Hence, the specific results in this subsection are about whether products, defined by their vectors of co-presence with industries, reveal clusters.

We implemented several clustering algorithms seeking stable clusters, on the matrix $J_{p,i}/(u_p u_i)$. Some of the algorithms implemented were: k-means, affinity propagation, spectral clustering, and DBSCAN (or density-based clustering), which are among the most widely used approaches. We expect two features from the exercise of finding clusters. First and most importantly, different algorithms should return approximately the same clusters. Second (and less importantly), the clusters should approximately overlap with the clus-

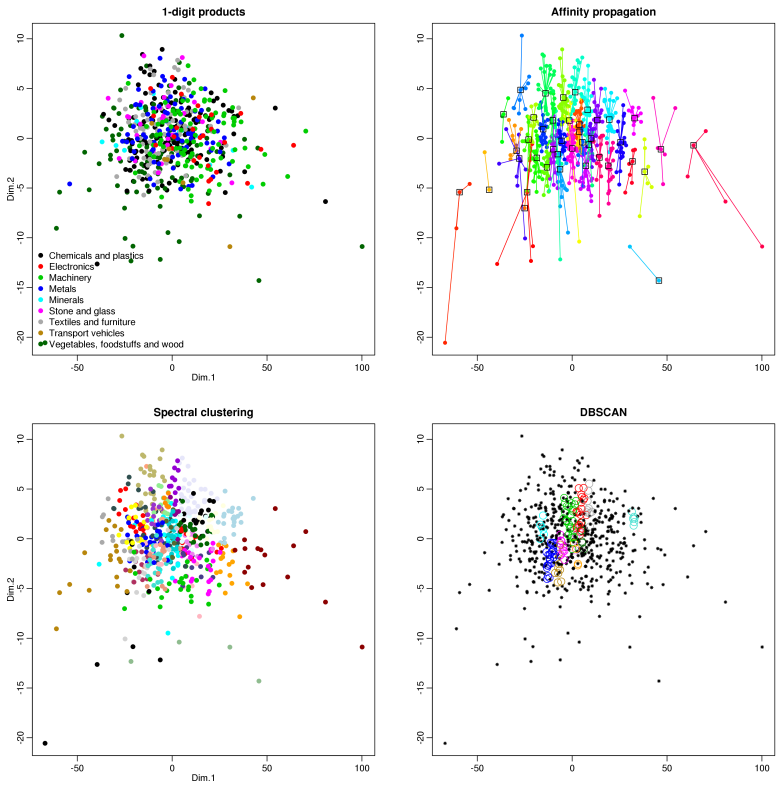

Figure 14: **Top-Left:** Visualization of similarities between products given by how they co-occur with industries, and colors depicting products that belong to the same 1-digit category. **Top-Right:** Affinity propagation clustering algorithm, with 41 clusters. **Bottom-Left:** Spectral clustering, where we have set the algorithm to seek 41 clusters. **Bottom-right:** Density-based clustering (DBSCAN), which found a maximum of 11 clusters here depicted. Black markers are cluster-less products that the algorithm allows. As can be seen, the algorithms are not stable, and do not display any correlation with the natural clustering from the classification.

tering imposed by the classification scheme at higher orders of aggregation. None of these two conditions were satisfied in our exercise. Just to show an example, we present in Figure 14 some of the clusters.

As is observed in fig. 14, the data does not exhibit well-defined clusters (or, at least, significant efforts should be devoted to exploring this particular question in more detail). Said differently, when we look at the linkages that cross between products and industries, linkages do not form clusters. This does not mean that industries do not serve as the pool of know-how that is combined to generate exports. It only means that the way in which industries provide that know-how is distributed across products. Thus, fig. 14 suggests that products seem to use the whole diversity of industrial resources available in a place. As of now, these results seem to cast doubts on the common use of "industrial clusters" in the analysis of export competitiveness (or, at least, on the literal interpretation of the notion of "cluster" in this context).

We conclude this subsection by saying that the joint Product-Industry Space would not make sense, based on the above discussion, at least not with the information we currently have. This does not invalidate the aim of the project as the presence or absence of industries may still determine the export possibilities in cities. In fact, it provides support for the construction of aggregate measures that capture the overall availability of ingredients required to export products. In the next section, we construct such measures.

4.5 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:[11]

$$\underbrace{\mathrm{D}(t)}_{c\times p} = \underbrace{\mathscr{C}(t)}_{c\times a} \cdot \underbrace{\mathscr{P}(t)}_{a\times p}. \tag{15}$$

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 (15) 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 equation (15).

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

In this first definition we will interpret equation (15) in the following ways:

(a) We will consider the industries to be the ingredients that drive exports, so i will denote the index of industries.

[11] 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.

(b) $\mathscr{P}$ 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.

(c) 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

$$D_{c,p}^{(1)} = \sum_i modRCA_{c,i} \left(\frac{\Pr_{(worker)}(i|p)}{\sum_{i'} \Pr_{(worker)}(i'|p)} \right), \tag{16}$$

where

$$\Pr_{(worker)}(i|p) = \frac{E_{p,i}/E_{tot}}{E_p/E_{tot}}. \tag{17}$$

4.5.2 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.

(b) $\mathscr{P}$ 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.

(c) 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

$$D_{c,p}^{(2)} = \sum_i modRCA_{c,i} \left(\frac{\Pr_{(city)}(i|p)}{\sum_{i'} \Pr_{(city)}(i'|p)} \right), \tag{18}$$

where

$$\Pr_{(city)}(i|p) = \frac{J_{p,i}/N_c}{u_p/N_c}, \tag{19}$$

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 (14) but with the matrices of *modRCA*. [12]

[12] Sometimes it is useful to express everything in terms of matrices and products of matrices. Let $X_{(c,i)}$ be the matrix of industry

4.5.3 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 equation (15). 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).

(b) $\mathscr{P}$ will be a *product* $\times$ *product* similarity matrix, based on a simple Pearson correlation between the rows of another *product* $\times$ *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.

(c) 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

$$D^{(3)}_{c,p} = \sum_{p' \neq p} modRCA_{c,p'} \left(\frac{\text{cor}(\mathbf{p}', \mathbf{p})}{\sum_{p'' \neq p} \text{cor}(\mathbf{p}'', \mathbf{p})} \right), \tag{22}$$

where $\mathbf{p}$ are the rows of the matrix defined by the elements

$$\frac{approxlog\left(\frac{J_{p,i}}{J_{p,p}J_{i,i}}, 500\right) - approxlog(0,500)}{-approxlog(0,500)}, \tag{23}$$

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 (23) 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.

---

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

$$\begin{aligned} [\mathbf{I}]_i &= \sum_c [M_{(c,i)}]_c \\ [\mathbf{P}]_p &= \sum_c [M_{(c,p)}]_c. \end{aligned} \tag{20}$$

Given this, as well,

$$\begin{aligned} \hat{\Pr}\{i|p\} &= \frac{[J]_{p,i}/N_c}{[\mathbf{P}]_p/N_c} \\ &= N_c \times \left[D^{-1}_{(p)} \cdot M^T_{(c,p)} \cdot M_{(c,i)} \right]_{p,i}, \end{aligned} \tag{21}$$

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

4.5.4 Density #4: $D^{(4)}_{c,p}$

Finally, our fourth definition of density is practically identical to the third definition in Equation (22). But to see the difference, note that in equation (22) 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).

(b) $\mathscr{P}$ will be a $product \times product$ similarity matrix, based on a simple Pearson correlation between the rows of another $product \times industry$ 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.

(c) 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

$$D^{(4)}_{c,p} = \sum_{p' \neq p} \frac{modRCA_{c,p'}}{\sum_{p'' \neq p} modRCA_{c,p''}} \, \text{cor}(\mathbf{p}', \mathbf{p}), \tag{24}$$

where $\mathbf{p}$ are the rows of the matrix defined by the elements

$$\frac{approxlog\left(\frac{J_{p,i}}{u_p u_i}, 500\right) - approxlog(0,500)}{-approxlog(0,500)}, \tag{25}$$

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 (25) is simply a statement of whether product p and industry i co-occur in cities more frequently than what is expected.

4.6 Empirical results

Before we present our empirical results, it is important to state in words the interpretation of each of our densities, Equations (16), (18), (22) and (24), when making a reference to a specific city c and a specific product p:

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

$D^{(2)}_{c,p}$**:** 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.

$D^{(3)}_{c,p}$**:** 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.

$D^{(4)}_{c,p}$**:** 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.

4.6.1 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 12 we report the pair-wise correlations between the density variables. As expected, all densities are positively correlated, yet they are not perfect

Table 12: Pairwise correlations between density variables.

	$D^{(1)}_{c,p}$	$D^{(2)}_{c,p}$	$D^{(3)}_{c,p}$	$D^{(4)}_{c,p}$
$D^{(1)}_{c,p}$	1	0.723	0.494	0.289
$D^{(2)}_{c,p}$	0.723	1	0.676	0.386
$D^{(3)}_{c,p}$	0.494	0.676	1	0.215
$D^{(4)}_{c,p}$	0.289	0.386	0.215	1

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".[13] The specifications are all the possible combinations defined by the following pieces:

$$\Delta Y_{t,t+\Delta t} = \beta_0 + \beta_1 Y_t + \mathbf{D_t}\boldsymbol{\beta} + FE, \quad \text{for each city } c \text{ and product } p. \tag{26}$$

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.

Figure 15 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.

[13] See `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

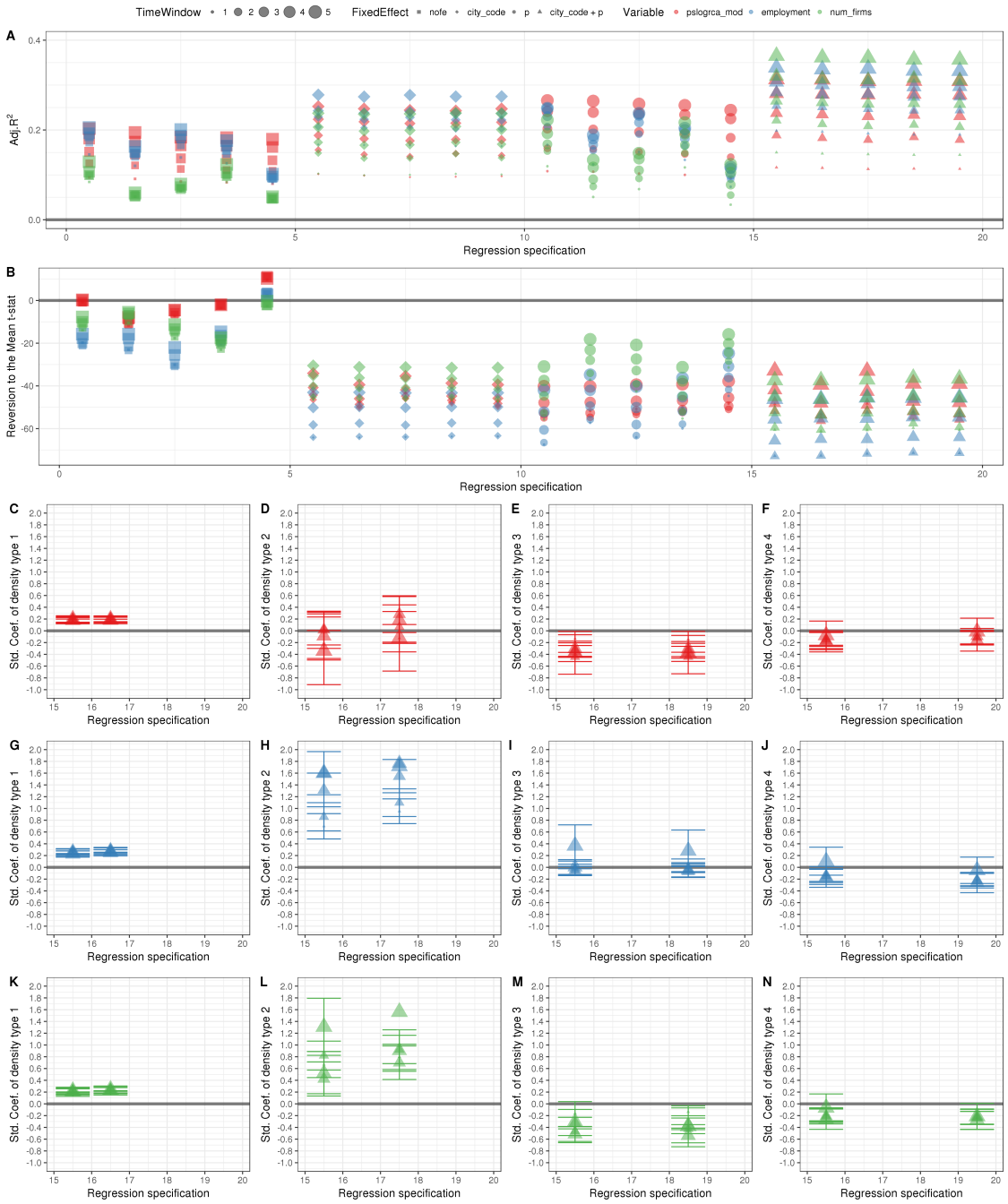

Figure 15: Visualizations of features of the results of 300 regressions (each dot refers to one of the regressions). **Panel A**: Adjusted R^2 of the density regressions. **Panel B**: t-statistic of the term for the reversion to the mean. **Panels C-F**: Estimated coefficients (standardized), with 95% 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.

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 fig. 15, 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 12. 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. Tables 13 to 15 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 15 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).

Table 13: **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.

	Dependent variable:			
	Change in modRCA in 5 years			
	(1)	(2)	(3)	(4)
modRCA initial year	-0.541^{***}	-0.537^{***}	-0.539^{***}	-0.534^{***}
	(0.014)	(0.016)	(0.014)	(0.016)
$D^{(2)}$		-0.122		-0.176
		(0.287)		(0.288)
$D^{(3)}$			-0.370^{**}	-0.381^{**}
			(0.184)	(0.184)
City fixed effects	YES	YES	YES	YES
Product fixed effects	YES	YES	YES	YES
Observations	6,595	6,595	6,595	6,595
R^2	0.378	0.378	0.379	0.379
Adjusted R^2	0.309	0.309	0.309	0.309
Residual Std. Error	0.831 (df = 5932)	0.831 (df = 5931)	0.831 (df = 5931)	0.831 (df = 5930)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 14: **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.

	Dependent variable:			
	Change in employment in 5 years			
	(1)	(2)	(3)	(4)
Employment initial year	-0.840^{***}	-0.882^{***}	-0.840^{***}	-0.884^{***}
	(0.018)	(0.019)	(0.018)	(0.019)
$D^{(2)}$		1.717^{***}		1.742^{***}
		(0.251)		(0.252)
$D^{(3)}$			0.280	0.346^{*}
			(0.180)	(0.180)
City fixed effects	YES	YES	YES	YES
Product fixed effects	YES	YES	YES	YES
Observations	6,595	6,595	6,595	6,595
R^2	0.399	0.403	0.399	0.404
Adjusted R^2	0.331	0.337	0.332	0.337
Residual Std. Error	0.818 (df = 5932)	0.815 (df = 5931)	0.818 (df = 5931)	0.814 (df = 5930)
Note:				*p<0.1; **p<0.05; ***p<0.01

4.6.2 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

Table 15: **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.

	Dependent variable:			
	Change in number of firms in 5 years			
	(1)	(2)	(3)	(4)
Number of firms initial year	-0.631^{***}	-0.661^{***}	-0.628^{***}	-0.658^{***}
	(0.017)	(0.018)	(0.017)	(0.018)
$D^{(2)}$		1.561^{***}		1.534^{***}
		(0.242)		(0.243)
$D^{(3)}$			-0.382^{**}	-0.310^{*}
			(0.177)	(0.177)
City fixed effects	YES	YES	YES	YES
Product fixed effects	YES	YES	YES	YES
Observations	6,595	6,595	6,595	6,595
R^2	0.421	0.425	0.421	0.425
Adjusted R^2	0.356	0.361	0.357	0.361
Residual Std. Error	0.802 (df = 5932)	0.800 (df = 5931)	0.802 (df = 5931)	0.799 (df = 5930)
Note:				*p<0.1; **p<0.05; ***p<0.01

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")?[14]

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):

$$\textit{Absolute absence at time } t \quad \longrightarrow \quad \textit{At least a firm with at least 1 effective employee at time } t+\Delta t$$

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

$$A_{t,t+\Delta t} = \text{logit}(\beta_0 + \mathbf{D_t}\boldsymbol{\beta} + FE). \tag{27}$$

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

[14]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.

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.

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 16 and 17, 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)[15] In other words, the 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

[15]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.

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, howerver, must be understood differently from AUC. We will compute the AIC based on the training set, and the AUC on the test set.

Table 16: 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.

	Logistic regression for appearances			Quality of statistical model		p-value for density coefficient			
#	Time period	Densities	Observations	AIC (training set)	Area Under the Curve (test set)	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$	$D^{(4)}$
1	5	All	37,758	5,870.7	0.827	0.0001	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$
2	5	Single	37,758	5,934.1	0.822		$\leq 10^{-5}$		
3	5	Single	37,758	6,215.4	0.786			$\leq 10^{-5}$	
4	5	Single	37,758	6,231.3	0.765	$\leq 10^{-5}$			
5	5	Single	37,758	6,532.1	0.714				$\leq 10^{-5}$
6	4	All	75,330	11,953.5	0.828	0.001	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$
7	4	Single	75,330	12,056.8	0.817		$\leq 10^{-5}$		
8	4	Single	75,330	12,651.4	0.783			$\leq 10^{-5}$	
9	4	Single	75,330	12,710.4	0.752	$\leq 10^{-5}$			
10	4	Single	75,330	13,160.5	0.729				$\leq 10^{-5}$
11	3	All	112,592	16,644.7	0.830	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$
12	3	Single	112,592	16,797.1	0.817		$\leq 10^{-5}$		
13	3	Single	112,592	17,570.4	0.750	$\leq 10^{-5}$			
14	3	Single	112,592	17,652.2	0.806			$\leq 10^{-5}$	
15	3	Single	112,592	18,334.9	0.738				$\leq 10^{-5}$
16	2	All	150,350	20,571.3	0.828	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	0.0001
17	2	Single	150,350	20,774.1	0.824		$\leq 10^{-5}$		
18	2	Single	150,350	21,696.3	0.760	$\leq 10^{-5}$			
19	2	Single	150,350	21,809.4	0.782			$\leq 10^{-5}$	
20	2	Single	150,350	22,750.0	0.773				$\leq 10^{-5}$
21	1	All	188,356	23,151.9	0.830	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$
22	1	Single	188,356	23,370.9	0.818		$\leq 10^{-5}$		
23	1	Single	188,356	24,403.4	0.761	$\leq 10^{-5}$			
24	1	Single	188,356	24,483.0	0.802			$\leq 10^{-5}$	
25	1	Single	188,356	25,303.2	0.763				$\leq 10^{-5}$

The first conclusion we can draw from tables 16 and 17 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 16 the corresponding ROC curve for the 1-year-period appearance model with

Table 17: Same results as Table 16 but showing the z-statistics of the coefficients for the densities. All are positive and large (i.e., statistically significant).

	Logistic regression for appearances			Quality of statistical model		z-statistic for density coefficient			
#	Time period	Densities	Observations	AIC	Area Under the Curve	$D^{(1)}$	$D^{(2)}$	$D^{(3)}$	$D^{(4)}$
1	5	All	37,758	5,870.7	0.827	3.8	10.7	5.5	4.1
2	5	Single	37,758	5,934.1	0.822		23.0		
3	5	Single	37,758	6,215.4	0.786			19.5	
4	5	Single	37,758	6,231.3	0.765	19.0			
5	5	Single	37,758	6,532.1	0.714				8.3
6	4	All	75,330	11,953.5	0.828	3.3	16.3	7.4	5.6
7	4	Single	75,330	12,056.8	0.817		32.1		
8	4	Single	75,330	12,651.4	0.783			26.2	
9	4	Single	75,330	12,710.4	0.752	25.3			
10	4	Single	75,330	13,160.5	0.729				12.6
11	3	All	112,592	16,644.7	0.830	6.3	18.3	9.1	5.0
12	3	Single	112,592	16,797.1	0.817		37.4		
13	3	Single	112,592	17,570.4	0.750	30.6			
14	3	Single	112,592	17,652.2	0.806			30.1	
15	3	Single	112,592	18,334.9	0.738				14.8
16	2	All	150,350	20,571.3	0.828	7.4	20.5	11.0	4.0
17	2	Single	150,350	20,774.1	0.824		41.5		
18	2	Single	150,350	21,696.3	0.760	34.1			
19	2	Single	150,350	21,809.4	0.782			33.6	
20	2	Single	150,350	22,750.0	0.773				15.7
21	1	All	188,356	23,151.9	0.830	5.7	21.3	11.4	6.5
22	1	Single	188,356	23,370.9	0.818		42.7		
23	1	Single	188,356	24,403.4	0.761	34.2			
24	1	Single	188,356	24,483.0	0.802			36.1	
25	1	Single	188,356	25,303.2	0.763				16.6

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. 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 18. The confusion matrix shows that there were 479

Table 18: 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".

	Actual - Not Appearance	Actual - Appearance
Predicted - Not Appearance	$TN = 28,240$	$FN = 88$
Predicted - Appearance	$FP = 9,163$	$TP = 391$

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.

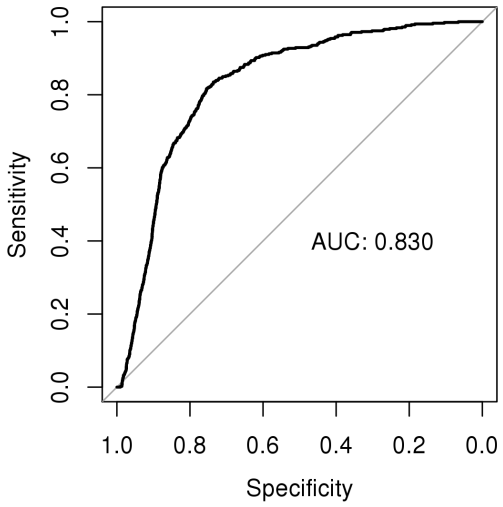

Figure 16: ROC curve over test set for predicting product appearances in cities from 2013 to 2014, having fitted a logistic model for all previous 1 year transition periods.

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.

4.7 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 3 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 2). 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 (figs. 2 to 5). This suggests there are factors that induce growth **multiplicatively**.
- The data suggests that **product diversification** is the main driver of firm growth (table 5 and fig. 7).
- 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 (fig. 8). 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 fig. 13, table 6, and equation (11). 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 fig. 14, 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 (fig. 15). 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 14 and 15).
- 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 16), 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). In the next section we will jump into applying Machine Learning techniques to boost our predictive power using the full disaggregated information about industrial presences in cities. There, we will revise again the concepts related to ROC curves, AUC, train vs. test sets, and other notions related to Machine Learning.

5 Machine learning methods

In the previous section, we explored the concept of 'similarity' as co-occurrence in various domains which suggested that the appearance of new products may be somehow determined by the industrial composition of places. After discovering the *mechanisms* that drive diversification of product exports, in this section, we *predict* appearances of new exports in cities. By utilizing Machine Learning techniques and methodologies, we are able to make stronger and more robust predictions.

Thus, we ask a slightly different question: Given *all* the *specific* industries in a city and the densities created in the previous sections, what is the probability of it exporting a certain product? Further, can we predict growth in these product exports using this data? Finally, how well can we predict the emergence of new export sectors in these cities?

5.1 Why machine learning?

Traditional regression methods such as generalized least squares or maximum likelihood estimators fail us in two major ways.

- Over-fitting: For each product, we have 62 cities or observation with approximately 400 possible industries and 4 densities forming 400+ explanatory variables. Since the total number of explanatory variables is much larger than the number of observations, traditional methods such as ordinary least squares would lead to "over-fitting"; they would perfectly fit the data.
- Functional form: Tradition regression methods require some prior knowledge of the functional form of the relationship between industries and products. We have no such prior on how industries interact to enable a city to export a product though we suspect that some interaction effects may exist. Exploring the infinite space of possible interaction effects is not trivial.

Machine learning (ML) gets around the problem of over-fitting by using regularizers[16]. These penalize complexity and converge on simpler models that are less likely to overfit. Some machine learning algorithms like decision trees[17] and support vector machines, do not require a functional form to be specified.

In general, tuned machine learning algorithms have shown to perform better than traditional econometric techniques at prediction[18]. The downside is that ML models can be difficult to unpack or interpret. The coefficients returned cannot be interpreted a correlations in the traditional econometric sense. This is made even worse by ensemble methods. Therefore ML algorithms provide strong predictive power but weak explanatory power.

[16] This adds an regularization parameter that needs to be tuned but techniques, such as k-fold cross-validation, exist to tune this parameters

[17] it can be argued that decision trees impose their own functional form. This is true but the hierarchical nature allows for a lot more flexibility. The use of ensemble methods like random forests provide even greater flexibility

[18] *Prediction* is accurately identifying the $\hat{y}$ whereas *estimation* is accurately predicting the β. ML algorithms sacrifice un-biasedness for low variance.

5.2 Machine learning algorithms

As part of this project, we explore multiple machine learning techniques. These are:

- LASSO: LASSO adds a *regularizer* term to the standard least squared models.

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|^2 + \lambda \|\beta\| \right\} \tag{28}$$

  where $\lambda\|\beta\|$ is the L1 regularizer which selects only the subset of the independent variables (X) that have high explanatory power. The model is useful if we believe the true model to be sparse i.e. only a few of the independent variables explain the dependant variable.

- Ridge: Instead if we believe the true model to be dense i.e. almost all independent variables explain the dependant variables, we may choose the Ridge L2 regularizer instead:

$$\min_{\beta} \left\{ \frac{1}{N} \|y - X\beta\|^2 + \lambda \beta^2 \right\} \tag{29}$$

- Random Forest: We can also train a decision tree to predict probability of exports. But deep trees i.e. with a lot of independent variables, tend to overfit the training set. Random Forests (RF) overcomes this with bootstrapping. It creates a number of decision trees, each trained on a bootstrapped random sample, and averages their predictions.

- Support Vector Machines: Given a training set, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where y_i can be -1 or 1, linear Support Vector Machines (SVM) find the hyperplane that maximally divides the groups of points for which $y_i = 1$ and from those with $y_i = -1$. Figure 17 shows an example of one such hyperplane. By utilizing higher order and non-polynomial kernels, we are able to fit a hyperplane in a tranformed (usually higher order) feature space.

- Gradient Boosted Trees: Gradient Boosted Trees (GBT), like RF, is another ensemble method that combines multiple decision trees. But unlike Random Forests, GBT builds the model in stages by fitting a tress to the *pseudo-residuals*. XGBoost is a popular implementation of GBT that we utilize in this paper.

5.3 Defining metrics

A test of a prediction algorithm is the percentage of out-of-sample (i.e. data it has never encountered before) records it is able to classify or predict correctly. We do two machine learning *regressions*: (1) predicting the RCA of product exports in a city (2) predicting the *change* in RCA of product exports in a city. We also do one machine learning *classification* where we predict the *appearance* of new export products. This section defines how we measure accuracy for these tests.

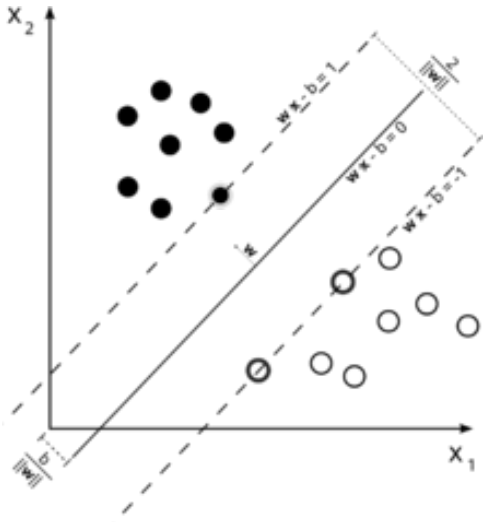

Figure 17: An example of a linear hyperplane. Courtesy: wikimedia/Public

For the regressions, we use R^2 to determine goodness of fit. Higher values would indicate that we are able to more accurately predict the outcome variable. High R^2 values are tough to achieve especially is out-of-sample test data since there may be a number of factors not captured by the data. In addition, there may be measurement error making the outcome variable noisier. We would consider a model test R^2 or around 0.2 or higher to be a good predictor.

A more nuanced metric exists for classification problems. As we noted previously, most cities will export only a few products and appearance of new products is a rare phenomenon. Therefore, a classification algorithm that classifies all products as not appearing will do very well. As an example, if only 5 out of 100 cities export product A, an algorithm that simply says that no city will export product A would have an accuracy of 95%. Therefore, we need to look at the *types* of errors made by the algorithm to understand performance.

A classification algorithm can make two types of errors:

- False positive (FP): Predicts that a product should be exported when it isn't.
- False negative (FN): Predicts that a product will not be exported when it is.

For completeness, we also define the following:

- True positive (TP): Correctly predicts that a product would be exported.
- True negative (TN): Correctly predicts that a product will not be exported.

One way to express the results from the classification is as a **confusion matrix** as shown in fig 18. Note that this allows you to see types of errors made.

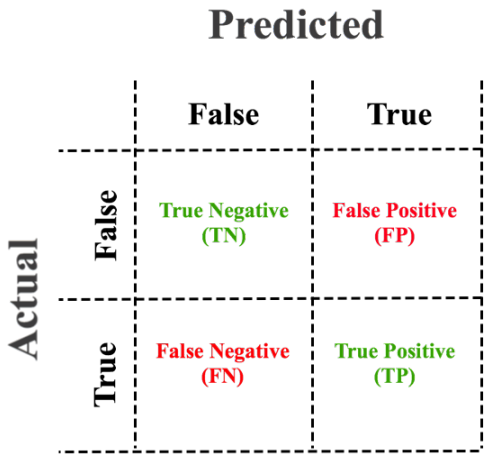

Figure 18: A confusion matrix

There is a trade-off between these errors. We may be able to increase the true positive rate at the cost of higher number of false positives. This relationship is captured in the **receiver operating characteristic (ROC) curve**. At each point of the curve, a new confusion matrix can be created. Another metric for the performance of the classifier is the **area under the curve (AUC)** of the ROC. A perfect classifier that makes no errors would have an AUC of 1 while a classifier that randomly guesses would have an AUC of 0.5.

5.4 Methodology

5.4.1 Specification

In the previous section, we looked three ways of calculating RCA for an export item - based on export value, employment at firms, or number of firms. Here, we predict the levels and changes in all of these metrics of RCA. When predicting levels, the three specifications are identical except for the outcome variable:

$$\begin{aligned} modRCA_{cpt} &= f(D^{(1)}_{cpt}, D^{(2)}_{cpt}, D^{(3)}_{cpt}, D^{(4)}_{cpt}, modRCA_{cit}, ..., modRCA_{cIt}, \gamma_c, \alpha_p, t) \\ logemp_{cpt} &= f(D^{(1)}_{cpt}, D^{(2)}_{cpt}, D^{(3)}_{cpt}, D^{(4)}_{cpt}, modRCA_{cit}, ..., modRCA_{cIt}, \gamma_c, \alpha_p, t) \\ logfirms_{cpt} &= f(D^{(1)}_{cpt}, D^{(2)}_{cpt}, D^{(3)}_{cpt}, D^{(4)}_{cpt}, modRCA_{cit}, ..., modRCA_{cIt}, \gamma_c, \alpha_p, t) \end{aligned} \tag{30}$$

where:

$$\begin{aligned} \gamma_c \text{ are dummies for city} \\ \alpha_p \text{ are dummies for product} \end{aligned} \tag{31}$$

When predicting changes, we include the level values for the base year as well. For a given window, $w \in (1,5)$:

$$\begin{aligned}
\Delta modRCA_{c,p}(t,t+w) &= f(modRCA_{c,p}(t), D^{(1)}_{c,p}(t), D^{(2)}_{c,p}(t), D^{(3)}_{c,p}(t), D^{4}_{c,p}(t), modRCA_{c,i}(t), ..., modRCA_{c,I}(t), \gamma_c, \alpha_p) \\
\Delta logemp_{c,p}(t,t+w) &= f(logemp_{c,p}(t), D^{(1)}_{c,p}(t), D^{(2)}_{c,p}(t), D^{(3)}_{c,p}(t), D^{4}_{c,p}(t), modRCA_{c,i}(t), ..., modRCA_{c,I}(t), \gamma_c, \alpha_p) \\
\Delta logfirms_{c,p}(t,t+w) &= f(logfirms_{c,p}(t), D^{(1)}_{c,p}(t), D^{(2)}_{c,p}(t), D^{(3)}_{c,p}(t), D^{4}_{c,p}(t), modRCA_{c,i}(t), ..., modRCA_{c,I}(t), \gamma_c, \alpha_p)
\end{aligned} \tag{32}$$

Similarly, when predicting appearance, we train the algorithms on the following specification:

$$A_{c,p}(t,t+w) = f(D^{(1)}_{c,p}(t), D^{(2)}_{c,p}(t), D^{(3)}_{c,p}(t), D^{4}_{c,p}(t), modRCA_{c,i}(t), ..., modRCA_{c,I}(t), \gamma_c, \alpha_p) \tag{33}$$

We defined 'appearance' as having absolute absence at time t to at least a firm with at least 1 effective employee at time $t+w$. Therefore, all three outcome measures - export value, employment, and firms - are zero at time t.

5.4.2 Other methodological details

All results presented are on an out-of-sample test set. When predicting levels, we take a random 20% sample as the test set. For the other two models, we take the last time period as the test set. All of the models use k-fold cross-validation (or grid-search when there are multiple parameters) to tune the hyper-parameters of the model. As we have seen, the data for classification model is highly unbalanced; new product appearance is rare. A number of solution exist to tackle this imbalance. We choose to balance the dataset by giving higher weight to records with new product appearance[19]. This allows us to reduce the number of false negatives in the prediction.

5.5 Results

Here we present the results from the two regressions models and the one classification model.

5.5.1 Predicting levels

Figure 19 shows the performance of each of the models for the three specifications. Levels tend to be easier to predict since ubiquity and diversity are themselves strong predictors. Further a level might not change substantially between periods. We see that SVM performs the best when predicting export value or number of firms with an R^2 of greater than 0.6. Xgboost, an implementation of GBT, performs best out of the models and predicts employment with a comparitively lower R^2 of 0.23.

[19] Weights are chosen such that the sum of all weights of records where products appear is equal to the sum of weights where no new products appear.

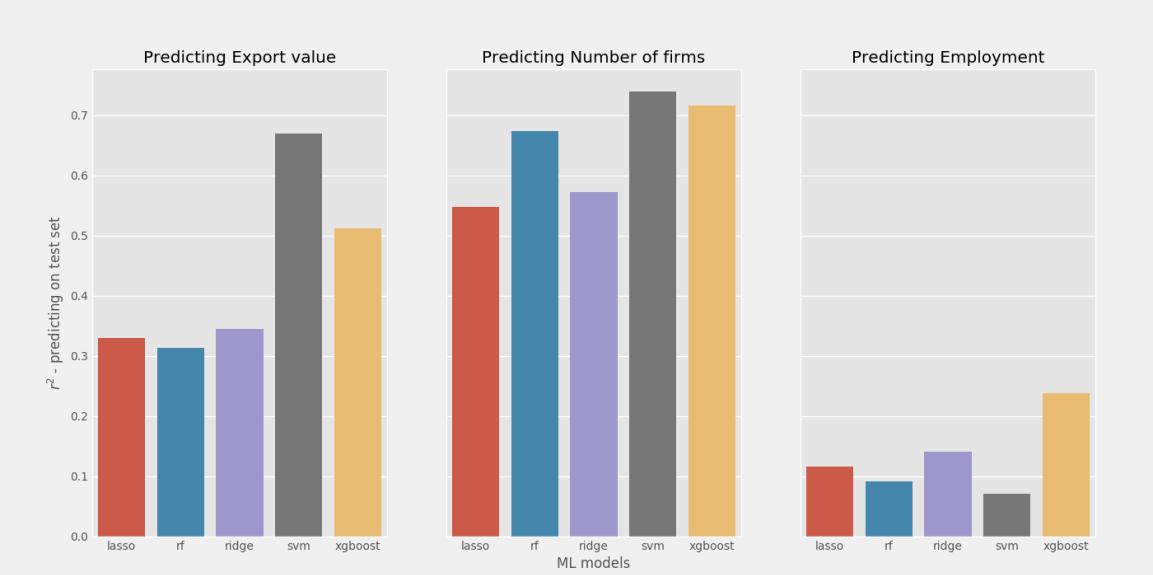

Figure 19: Predicting levels of export variables across cities

5.5.2 Predicting differences

Predicting changes in product RCA is a lot more difficult. We again train the five types of algorithms for each outcomes and repeat for each change window. The results are shown in Figure 20. Xgboost outperforms most of the other models in all specifications with R^2 values ranging from around 0.17 to almost 0.3 on the test set predictions. Note that it is easier to predict changes over a longer time horizon even though there are fewer records on which to train the model. Part of the explanation may be that there is a general growth trend impacting all cities and, since the RCAs are relatives to the world, *all* RCAs increase making it easier to predict the change. Another reason could be that change induced by the presense of industries takes a few years to be observable. In shorter time windows, the magnitude of the change is smaller and random fluctuations make it a very noisy metric.

5.5.3 Predicting appearances

Predicting changes, as we did in the previous section, amounts to predicting growth in product export at the intensive margin i.e. how existing products grow (or shrink) in the presense of industries. Now, we consider the extensive margin; can we predict the appearance of new products in cities? We consider the city-product pairs that do not exist[20] and see if we can predict their appearance over varying time windows.

We use RF and GBT to predict these appearances. The results are shown in Figure 21. Note that in the previous section, we were able to achieve an AUC of around 0.83 using a logistic model. GBT significantly outperformed RF and the logistic model with an AUC of almost 0.9.

[20]These are pairs with an RCA of 0 - there are no firms producing this product in the city

Figure 20: Predicting changes of export variables across cities

Figure 22 shows the ROC curves for the two ML algorithms and the logistic regression from the previous section for different change windows[21]. Note that the all points on the logistic regression ROC curves (in green) are below those of GBT (in blue) indicating that GBT produces fewer errors. This makes a strong case for using ML for recommendations and predictions instead of traditional econometric techniques.

We can choose a threshold and present a confusion matrix showing the different errors. Here, we choose a threshold of 0.1 but this can be varied based on the acceptable trade-off between true positives and false positives. Recall that each point on the ROC curve represents one such trade-off point:

		Predicted	
		Doesn't Appear	Appears
Actual	Doesn't Appear	35905	1568
	Appears	334	447

Table 19: Confusion matrix with threshold 0.1

Note that the even the gradient boosted trees model is hardly infallible. It does make some classification errors and fails to identify some of the appearance. Though, given how rare these are, it does remarkably well by identifying more than half of them. This true positive rate can be increased by choosing a lower threshold but it also results in a higher false positive rate. This can be seen in Table 20 which uses a threshold of 0.02.

[21] Figure 21 and 22 are from different runs and hence report slightly different numbers.

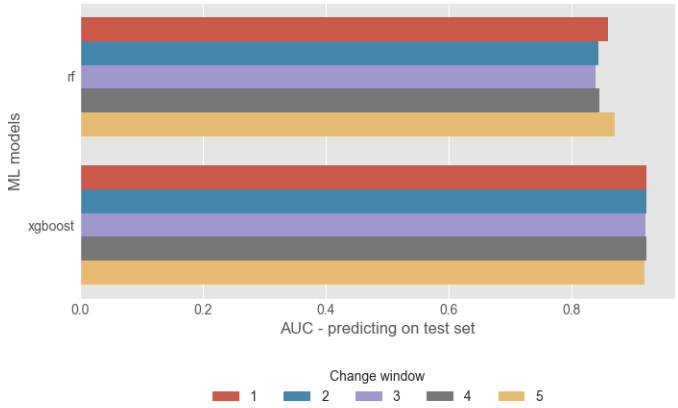

Figure 21: Predicting appearance of new product exports at city level

		Predicted	
		Doesn't Appear	Appears
Actual	Doesn't Appear	32500	4973
	Appears	136	645

Table 20: Confusion matrix with threshold 0.02

5.6 Discussion

GBT using a 5-year window was the best model when predicting changes in export variables and export appearances. Here, we use the trained models to predict intensive and extensive growth in export sectors. In particular, we look at discrepancies between predictions and reality. These predictions are not infallible; there may be legitimate reasons for why these discrepancies exist - measurement error, policy environment, security etc. But they lead to a deeper investigation into why these city-product pairs are different from all other. It may help us uncover factors constraining growth in these cities.

Some of these charts may be difficult to read due to the detail present. They can also be accessed online, along with other specifications, at `https://goo.gl/zxLSYF`.

5.6.1 Highest growth

Figure 23 shows the top 100 city-product pairs where the actual growth was lower than expected i.e. the difference between predicted growth and actual growth was the largest. Squares with lighter colors represent a smaller difference. Most of the differences are small though a couple of cities like Apartadó and Santa Marta have a few products with large gaps.

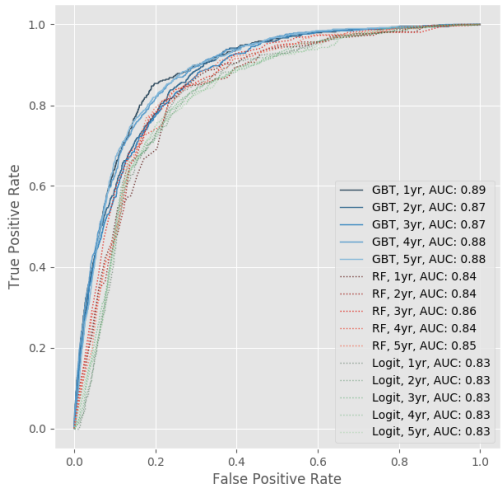

Figure 22: ROC/AUC of RF and GBT models

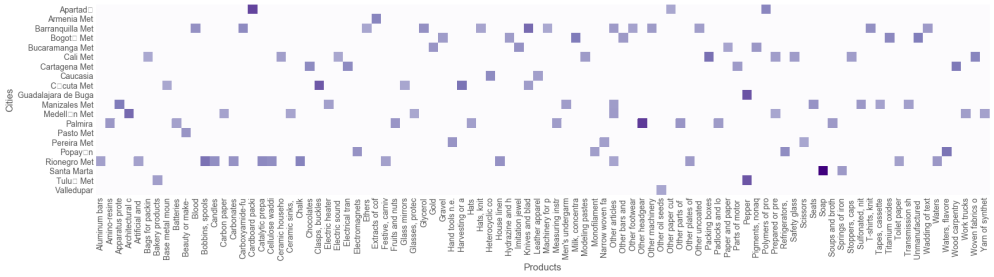

Figure 23: City-product pairs with the greatest differential between prediction and actual

5.6.2 Most likely appearances

Figure 24 shows the top 100 city-product pairs that were predicted to appear with the highest probability. It's interesting to note that most of these are in Yopal. It may be that Yopal is different in ways not captured by the data. But this should also prompt a deeper investigation into why Yopal does not export more products than it currently does. A diagnostic may reveal binding constraints that when relaxed would allow the city to expand its exports.

5.7 Conclusion

Machine learning algorithms are particularly suited for prediction tasks that use a wealth of data. In this section, we build on previous sections by utilizing the density matrices along with disaggregated industry data to predict the level, changes, and appearances of products. Using robust prediction models and methodolo-

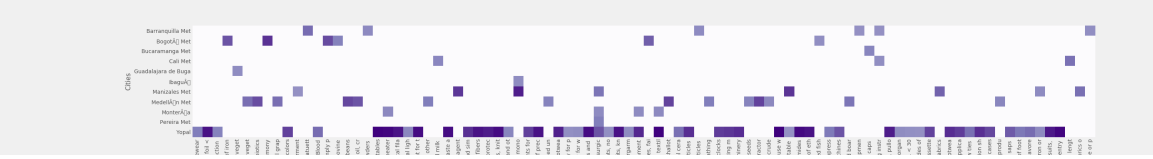

Figure 24: City-product pairs predicted to appear with the highest probability but absent from dataset

gies, we are able to identify future export growth areas with greater certainity. These city-export predictions could form a launching pad into deeper investigation into questions such as why these don't already exist, what actions can be taken to promote these, and what implications this has for economic growth in the city and in Colombia.

In this paper we used a number of techniques to understand the mechanisms of diversification and predict what these imply for growth in the export sector. While traditional econometric techniques can be used to test hypothesis and understand underlying mechnanisms, ts prediction may not be fit for use for normative guidance. Machine learning methodologies complement these traditional techniques through stronger and more stable predictions. These predictions have greater accuracy and hence are better suited to guide policy decisions.

Appendix A Dropped industries and products

Table A21: List of industries that were dropped for the regressions and machine learning analysis in this report. We were not able to find class names for some codes.

Code	Class name
0112	Cut flowers
0140	Farming services
0144	
1120	Service for oil and gas extraction
1571	Sugar
1589	Other food products n.e.c.
1702	
1710	Textile fibers
1810	Apparel, except fur
2239	Other service activities n.e.c.
2423	Pharmaceuticals
2424	Soaps and detergents, cleaning preparations
2529	Plastic articles n.e.c.
2899	Other fabricated metal products n.e.c.
3430	Parts and accessories for motor vehicles and their engines
3699	Other manufacturing n.e.c.
4290	
4530	Building of civil engineering works
4751	
5040	Sale, maintenance and repair of motorcycles and related parts and accessories
5051	Retail sale of automotive fuel
5124	Wholesale of livestock raw materials and live animals
5125	Wholesale of food, except coffee threshings
5127	Wholesale of beverages and snuff products
5135	Wholesale of pharmaceutical, cosmetic
5141	Wholesale of construction materials, hardware and glass
5190	Wholesale of different products n.e.c.
5211	Non-specialized stores food, beverages or tobacco
5219	Retail sale in non-specialized stores with food, beverages or tobacco predominating
5231	Retail sale of pharmaceutical, cosmetic
5233	Retail sale of apparel and accessories
5235	Retail sale of household appliances
5241	Retail sale of hardware, locks and glass
5249	Retail sale of other new consumer products n.e.c.
5269	Other retail sale not in stores
5819	
6211	National scheduled air transport of passengers
6599	Other financial intermediation n.e.c.
7411	Legal activities
7421	Architectural, engineering and other technical
7491	Labour recruitment and provision of personnel
7499	Other business activities n.e.c.
7512	Executive activities in public administration
7912	
8050	Higher education
8299	
8511	Hospital activities
8512	Medical practice activities
8519	Other human health activities
8522	
8532	Social work without accommodation
8544	
8551	
8610	
8699	
9191	Activities of religious organizations
9199	Activities of other membership organizations n.e.c.
9309	Other service activities n.e.c.
9500	Private households with employed persons

Table A22: List of products that were dropped for the regressions and machine learning analysis in this report. We were not able to find product names for some codes.

Code	Product name
0101	Horses
0105	Fowl
0203	Pork
0204	Lamb
0206	Edible offal
0208	Other meat
0210	Preserved meat
0301	Live Fish
0307	Molluscs
0401	Milk
0404	Whey
0407	Eggs, in shell
0408	Egg yolks
0409	Honey
0410	Edible animal products, n.e.c.

0502	Brushmaking hair
0504	Animal guts, except fish
0507	Ivory, tortoise-shell, whalebone, etc.
0508	Coral and shells
0601	Flower bulbs
0701	Potatoes
0702	Tomatoes
0704	Cabbages, cauliflower, kale, etc.
0705	Lettuce
0706	Carrots, turnips, beets, etc.
0707	Cucumbers
0801	Coconuts, Brazil nuts and cashews
0802	Other nuts
0807	Melons and papayas
0809	Apricots, cherries, peaches, plums
0812	Fruits and nuts, provisionally preserved
0814	Peel of citrus fruit or melons
0902	Tea
0906	Cinnamon
0907	Cloves
0909	Anise, fennel, etc.
1001	Wheat and meslin
1003	Barley
1004	Oats
1006	Rice
1007	Grain sorghum
1008	Buckwheat and other cereals
1103	Cereal groats, meals and pellets
1105	Potato flour
1107	Malt
1109	Wheat gluten
1201	Soya beans
1202	Peanuts
1204	Linseed
1205	Rape or colza seeds
1206	Sunflower seeds
1208	Flours of oil seeds
1209	Seeds used for sowing
1210	Hop cones, fresh or dried
1213	Cereal straw and husks
1214	Forage products
1301	Lac
1501	Pig and poultry fat, rendered
1503	Lard oil, tallow oil etc.
1504	Fats and oils of fish or marine mammals
1505	Wool grease
1506	Other animal fats and oils
1507	Soybean oil, crude
1509	Olive oil, virgin
1512	Sunflower-seed oil, crude
1514	Rapeseed, colza or mustard oil,
1521	Vegetable waxes and beeswax
1522	Degras and wax residues
1601	Sausages
1602	Other prepared or preserved meat
1603	Extracts and juices of meat or fish
1605	Prepared aquatic invertibrates
1703	Molasses
1802	Cocoa residues
1902	Pasta
1903	Tapioca
2002	Tomatoes, prepared or preserved
2003	Mushrooms, prepared or preserved
2004	Other vegetables, frozen
2204	Wine
2205	Wine, flavored
2206	Other fermented beverages
2207	Ethyl alcohol >80%
2301	Flours of meat or fish, unfit for human consumption
2302	Cereal residues
2303	Starch residues
2304	Solid soybean residues
2306	Solid vegetable oil and fat residues
2308	Vegetable materials used in animal feeding
2402	Cigars and cigarettes
2403	Other manufactured tobacco
2502	Unroasted iron pyrites
2503	Sulphur, crude

2506	Quartz
2510	Natural calcium phosphates
2511	Natural barium sulfate or carbonate
2512	Siliceous fossil meals and earths
2513	Pumice
2514	Slate
2515	Marble
2516	Granite, basalt etc.
2518	Dolomite
2519	Natural magnesium carbonate
2520	Gypsum
2521	Limstone
2522	Quicklime
2524	Asbestos
2528	Natural borates
2601	Iron ores and concentrates
2602	Manganese >47% by weight
2603	Copper ore
2604	Nickel ore
2606	Aluminum ore
2607	Lead ore
2608	Zinc ore
2609	Tin ore
2610	Chromium ore
2611	Tungsten ore
2614	Titanium ore
2615	Niobium, tntalum etc. ores
2616	Precious metal ores
2617	Other ores
2618	Granulated iron or steel slag
2619	Iron or steel slag
2620	Slag, ash and residues containing metals
2621	Other slag and ash
2702	Lignite
2705	Non-petroleum gases
2706	Tar distilled from coal, lignite etc.
2708	Pitch and pitch coke
2713	Petroleum coke
2715	Bituminous mixtures
2716	Electrical energy
2802	Sulfur, sublimed or precipitated
2805	Alkali metals, mercury etc.
2807	Sulfiric acid, oleum
2809	Phosphoric acid etc.
2810	Oxides of boron; boric acids
2812	Halides of nonmetals
2813	Sulfides of nonmetals
2814	Ammonia
2816	Hydroxides or peroxides of magnesium
2820	Manganese oxides
2822	Cobalt oxides and hydroxides
2824	Lead oxides
2826	Flourides
2829	Chlorates, bromates, y iodates
2831	Dithionites and sulfoxylates
2834	Nitrites, nitrates
2837	Cyanides
2840	Borates; peroxoborates
2841	Salts of oxometallic acids
2842	Other salts of acids
2844	Radioactive chemical elements
2845	Non-radioactive isotopes
2846	Compounds of rare-earth metals
2849	Carbides
2850	Hydrides, nitrides, azides, silicides and borides
2851	Inorganic compounds, liquid or compressed air
2901	Acyclic hydrocarbons
2903	Halogenated derivatives of hydrocarbons
2908	Derivatives of phenols
2910	Epoxides
2911	Acetals and hemiacetals
2913	Derivatives of aldehydes
2919	Phosporic esters
2925	Carboxyimide-function compounds
2926	Nitrile-function compounds
2927	Diazo-, azo-, or azoxy-compounds
2928	Organix derivatives of hydrazine
2929	Compounds with other nitrogen function

2935	Sulfonamides
2937	Hormones
2938	Glycosides
2940	Sugars, chemically pure, other than sucrose, lactose, maltose, glucose and fructose
2942	Other organic compounds
3001	Organs for therapeutic use
3201	Tanning extracts of vegetable origen
3501	Casein
3502	Albumins (water soluble proteins)
3602	Prepared explosives, except gunpowder
3603	Detonators
3604	Fireworks
3605	Matches
3606	Ferrocerium and other pyrophoric alloys
3702	Photographic film in rolls
3703	Photographic paper
3704	Photographic film, not developed
3705	Photographic film, developed
3706	Motion-picture film
3707	Chemical preparations for photographic uses
3803	Tall oil
3805	Turpentines
3807	Wood tar and oils
3813	Preparations for fire extinguishers
3817	Mixed alkylbenzenes
3818	Chemical elements for electronics
3819	Hydraulic fluids
3821	Prepared culture media for micro-organisms
3822	Diagnostic or laboratory reagents
3914	Ion-exchangers based on polymers
4001	Natural rubber
4003	Reclaimed rubber
4004	Scrap of rubber
4006	Other articles of unvulcanized rubber
4007	Vulcanized rubber thread and cord
4014	Rubber hygenic or pharmeceutical items
4102	Raw skins of sheep or lambs
4105	Tanned sheepskins
4106	Tanned skins of other animals
4108	Chamois leather
4109	Patent leather
4110	Waste of leather
4206	Articles of gut
4301	Other raw furskins
4302	Other tanned furskins
4303	Furskin apparel
4304	Artificial fur
4401	Fuel wood
4402	Wood charcoal
4406	Wooden railway ties
4408	Sheets for veneering for plywood
4412	Plywood
4413	Densified wood
4416	Casks, barrels, etc. of wood
4419	Wooden kitchenware
4502	Natural cork, debacked
4503	Articles of natural cork
4504	Agglomerated cork
4601	Products of plaiting materials
4701	Mechanical woodpulp
4702	Chemical woodpulp, disolving grade
4703	Chemical woodpulp, soda or sulfate
4704	Chemical woodpulp, sulfite
4705	Semichemical woodpulp
4706	Pulps of recovered paper fibers
4801	Newsprint
4806	Greaseproof paper
4807	Uncoated composite paper
4808	Corrugated paper and paperboard
4812	Filter blocks of paper pulp
4814	Wallpaper
4816	Other carbon paper
4904	Music, printed or in manuscript
4906	Original hand-drawn plans
4907	Unused stamps
5004	Silk yarn
5007	Woven silk fabrics
5101	Wool

5103	Wool or animal hair waste
5105	Wool or animal hair, combed
5107	Yarn of combed wool, not for retail sale
5109	Yarn of wool or animal hair, for retail sale
5110	Yarn of coarse animal hair or of horsehair
5111	Woven fabrics of carded wool
5112	Woven fabrics of combed wool
5202	Cotton waste
5203	Cotton, carded or combed
5205	Cotton yarn of >85%
5206	Cotton yarn of <85%
5212	Other woven cotton fabrics
5301	Flax, raw or processed
5303	Textile bast fibers
5305	Coconut and other vegetable textile fibers
5306	Flax yarn
5307	Yarn of textile bast fibers
5309	Woven fabrics of flax
5310	Woven fabrics of jute or of other textile bast fibers
5403	Artificial filament yarn
5404	Synthetic monofilament >67 dtex, thickness <1mm
5405	Artificial monofilament >67dtex t<1mm, strip, straws t<5mm
5406	Man-made filament yarn for retail sale
5408	Woven fabrics of artificial filament yarn
5501	Synthetic filament tow
5502	Artificial filament tow
5504	Artificial staple fibers, not processed for spinning
5505	Waste of man-made fibers
5506	Synthetic staple fibers, processed
5507	Artificial staple fibers, processed
5510	Yarn of artificial staple fibers, not for retail sale
5511	Yarn of man-made staple fibers, for retail sale
5512	Woven fabrics of >85% synthetic staple fibers
5513	Woven fabrics of <85% synthetic staple fibers
5514	Woven fabrics of <85% synthetic staple fibers mixed mainly with cotton <170 g/m2
5604	Rubber textiles
5605	Metallised yarn
5606	Gimp yarn
5608	Nets
5609	Articles of yarn, rope etc not elsewhere clasified
5701	Carpets, knotted
5702	Woven carpets and rugs
5703	Carpets, tufted
5704	Carpets of felt
5801	Woven pile fabrics
5803	Gauze
5805	Hand-woven tapestries
5809	Woven fabric incorporating metal threads
5810	Embroidery in the piece, in strips or in motifs
5811	Quilted textile products
5901	Textile fabrics coated with gum
5905	Textile wall coverings
5907	Other textile fabrics impregnated, coated or covered
5908	Textile wicks
5909	Textile hosepiping and similar tubing
5910	Transmission belts or belting, of textile material
6001	Pile fabrics, knit
6113	Garments knit with impregnated fibers
6116	Gloves, knit
6216	Gloves
6308	Needlecraft sets of woven fabric and yarn
6309	Used clothes and textiles
6501	Hat forms
6502	Hat shapes
6507	Headbands
6602	Walking sticks
6603	Parts of umbrellas or walking sticks
6701	Feathers or down
6702	Artificial flowers
6703	Human animal hair prepared for use in wigs
6704	Wigs
6801	Flagstones, of natural stone
6803	Worked slate
6807	Asphalt
6808	Panels of vegetable fibers
6809	Plaster articles
6814	Mica articles
6901	Bricks, blocks, and other ceramic goods

6906	Ceramic pipes
6909	Ceramic wares for laboratory, agriculture, or packing use
6914	Other ceramic articles
7001	Cullet and other scraps of glass
7002	Glass balls
7003	Glass, cast or rolled
7004	Drawn and blown glass
7006	Worked glass
7008	Multiple-walled insulating glass
7011	Glass envelopes
7014	Signaling glassware
7015	Clock or watch glasses and similar glasses
7018	Glass beads
7020	Other articles of glass
7101	Pearls
7102	Diamonds
7103	Precious stones
7104	Synthetic precious stones
7106	Silver
7109	Gold clad metals
7111	Platinum clad metals
7112	Scrap of precious metal
7114	Goldsmith and silversmith wares
7115	Other articles of precious metals
7116	Articles or pearls or precious stones
7118	Coin
7201	Pig iron
7203	Ferrous products from the reduction of iron ore
7205	Powders of iron or steel
7206	Iron and nonalloy steel
7207	Semifinished products of iron or nonalloy steel
7208	Flat-rolled iron, width >600mm, hot-rolled, not clad
7209	Flat-rolled iron, width >600mm, cold-rolled, not clad
7213	Hot rolled bars of iron
7214	Other bars of iron, not further worked than forged
7218	Stainless steel in ingots
7219	Flat-rolled products of stainless steel of a width >600 mm
7220	Flat-rolled products of stainless steel of a width <600 mm
7221	Bars of stainless steel, hot-rolled
7223	Wire of stainless steel
7224	Other alloy steel in primary form
7225	Flat-rolled products of other alloy steel, width >600 mm
7226	Flat-rolled products of other alloy steel, width <600 mm
7227	Bars of other alloy steel
7228	Other bars and rods of other alloy steel
7229	Wire of other alloy steel
7301	Sheet piling of iron or steel
7302	Railway construction material of iron or steel
7305	Other tubes and pipes, diameter >406.4 mm, of iron or steel
7322	Radiators for central heating of iron or steel
7401	Copper mattes
7402	Unrefined copper
7403	Refined copper and copper alloys
7405	Master alloys of copper
7408	Copper wire
7410	Copper foil <0.15 mm thick
7411	Copper tubes and pipes
7502	Nickel unwrought
7503	Nickel waste and scrap
7505	Nickel bars, wire etc.
7507	Nickel tubes and pipes
7508	Other articles of nickel
7601	Unwrought aluminum
7603	Aluminum powders
7611	Aluminum containers, >300 liters
7613	Aluminum containers for compressed or liquefied gas
7802	Lead waste or scrap
7804	Lead foil <2mm
7806	Other articles of lead
7901	Unwrought zinc
7902	Zinc waste and scrap
7903	Zinc powders
7904	Zinx bars and wire
7905	Zinc plates and foil
8001	Unwrought tin
8003	Tin bars and wire
8007	Other articles of tin
8101	Tungsten (wolfram)

8103	Tantalum
8104	Magnesium
8105	Cobalt
8108	Titanium
8109	Zirconium
8111	Manganese
8112	Other metals
8113	Cermets
8209	Articles for utensils, of cermet
8212	Razors
8304	Paper trays and similar office equipment, of base metal
8401	Nuclear reactors and related equipment
8403	Central heating boilers
8404	Auxiliary parts for use with boilers
8405	Water gas generators
8406	Steam turbines
8410	Hydraulic turbines, water wheels and regulators
8411	Gas turbines
8416	Furnace burners for liquid fuel
8420	Calendering or other rolling machines, other than for metals or glass
8426	Ships' derricks; cranes
8427	Fork-lift trucks
8435	Machines for wine and juice production
8440	Bookbinding machinery
8444	Machines to extrude, cut manmade textile fibres
8447	Knitting machines
8448	Auxiliary machinery for use with knitting and textile machines
8449	Machinery to manufacture felt
8450	Household- or laundry-type washing machines
8453	Machinery for preparing leather
8454	Machines used in metallurgy
8455	Metal-rolling mills
8456	Machines for working materials by laser and similar means
8457	Machining centers for working metal
8458	Lathes for removing metal
8459	Machine tools for drilling by removing metal
8461	Other machine tools for planing and cutting metals
8469	Word processing machines
8470	Calculating machines, cash registers etc.
8473	Parts and accessories for office machines
8476	Automatic goods-vending machines
8478	Machinery for preparing tobacco
8510	Electric shavers, hair clippers and hair-removing appliances
8513	Portable electric lamps
8517	Telephones
8518	Microphones
8519	Sound recording apparatus
8521	Video recording apparatus
8522	Parts and accessories for video or sound equipment
8525	Transmission apparatus for radio, telephone and TV
8526	Radar
8527	Reception apparatus for radio broadcasting
8528	Monitors and projectors
8529	Parts of radios, telephones, and T.V.s
8530	Electric signal and traffic controls
8532	Electrical capacitors
8533	Electrical resistors
8534	Electronic printed circuits
8540	Thermionic, cold cathode or photocathode tubes
8542	Electronic integrated circuits
8545	Carbon articles for eletrical purposes
8547	Insulating fittings for electrical machines
8602	Other rail locomotives
8603	Self-propelled railway coaches
8604	Railway service vehicles
8608	Railway track fixtures
8609	Containers for multimodal transportation
8701	Tractors
8702	Motor vehicles for the transport of >10 persons
8703	Cars
8704	Motor vehicles for transporting goods
8705	Special purpose motor vehicles
8706	Vehicle chassis fitted with engines
8707	Vehicle Bodies
8710	Tanks and other armored fighting vehicles
8712	Bicycles
8713	Carriages for disabled persons
8715	Baby carriages

8801	Gliders, hang gliders
8802	Other aircraft and spacecraft
8803	Parts of other aircraft
8804	Parachutes
8805	Aircraft launching gear
8902	Fishing vessels
8903	Pleasure or sport boats
8904	Tugs and pusher craft
8905	Floating or submersible drilling platforms
8906	Other vessels
8907	Other floating structures
9001	Optical fibers
9002	Lenses and other optical elements
9003	Frames for spectacles, goggles
9005	Binoculars amd other optical telescopes
9006	Photographic cameras
9007	Cinematographic cameras and projectors
9008	Still image projectors
9010	Apparatus and equipment for photographic laboratories, n.e.c.
9011	Optical microscopes
9012	Microscopes, other than optical
9013	Liquid crystal devices
9033	Other parts for machines and appliances
9101	Watches with cases of precious metal
9103	Clocks with watch movements
9104	Instrument panel clocks for vehicles
9106	Apparatus for measuring intervals of time
9108	Watch movements, complete
9109	Clock movements, complete
9110	Clock movements, complete, unassembled
9111	Watch cases and parts
9112	Clock cases
9113	Watch straps
9114	Other clock or watch parts
9201	Pianos
9202	Musical instruments, string
9205	Musical instruments, wind
9206	Musical instruments, percussion
9207	Musical instruments, electric
9209	Parts of musical instruments
9301	Military weapons, other than pistols
9302	Revolvers and pistols
9303	Other firearms
9304	Other arms (air guns, truncheons, etc.)
9305	Parts of military weapons
9306	Munitions of war
9307	Swords, cutlasses, etc.
9402	Medical, dental or veterinary furniture
9504	Articles for arcade, table or parlor games
9507	Fishing and hunting equipment
9508	Merry-go-rounds and other fairground amusements
9601	Worked ivory, tortoise-shell, etc.
9604	Hand sieves and riddles
9605	Travel sets
9611	Hand-operated stamps
9612	Typewriter ribbons and ink pads
9614	Smoking pipes
9701	Paintings and drawings
9702	Original engravings
9703	Sculptures
9705	Collectors pieces
9706	Antiques >100 years
9806	