Implied Comparative Advantage*

Ricardo Hausmann César A. Hidalgo Daniel P. Stock
Muhammed A. Yildirim†

January 2014

Abstract

Ricardian theories of production often take the comparative advantage of locations in different industries to be uncorrelated. They are seen as the outcome of the realization of a random extreme value distribution. These theories do not take a stance regarding the counterfactual or implied comparative advantage if the country does not make the product. Here, we find that industries in countries and cities tend to have a relative size that is systematically correlated with that of other industries. Industries also tend to have a relative size that is systematically correlated with the size of the industry in similar countries and cities. We illustrate this using export data for a large set of countries and for city-level data for the US, Chile and India. These stylized facts can be rationalized using a Ricardian framework where comparative advantage is correlated across technologically related industries. More interestingly, the deviations between actual industry intensity and the implied intensity obtained from that of related industries or related locations tend to be highly predictive of future industry growth, especially at horizons of a decade or more. This result holds both at the intensive as well as the extensive margin, indicating that future comparative advantage is already implied in todays pattern of production.

JEL Codes: F10, F11, F14, O41, O47, O50

*We thank Philippe Aghion, Pol Antràs, Sam Asher, Jesus Felipe, Elhanan Helpman, Asim Khwaja, Paul Novosad, Andrés Rodrìguez-Clare and Dani Rodrik for very useful comments on earlier drafts. We are indebted to Sam Asher and Paul Novosad for sharing the data on India and the Servicio de Impuestos Internos for sharing the data on Chile. All errors are ours.

†Hausmann: Center for International Development at Harvard University, Harvard Kennedy School and Santa Fe Institute. Hidalgo: The MIT Media Lab, Massachusetts Institute of Technology and Instituto de Sistemas Complejos de Valparaiso. Stock: Center for International Development at Harvard University and The MIT Media Lab. Yildirim: Center for International Development at Harvard University. Emails: ricardo_hausmann@harvard.edu (Hausmann), hidalgo@mit.edu (Hidalgo), daniel_stock@hks.harvard.edu (Stock), muhammed_yildirim@hks.harvard.edu (Yildirim).

Ricardian theory predicts that countries export the goods in which they have a comparative advantage, meaning that they enjoy a higher relative productivity. Although Ricardo introduced this idea almost two centuries ago (Ricardo, 1817), the multi-country multi-product version of his model has only recently been formalized and subjected to rigorous empirical testing (Eaton and Kortum, 2002; Costinot et al., 2012). These models infer a country's productivity in a certain industry from its observed pattern of trade and have been successful in explaining a significant portion of bilateral trade. Yet, these models can only infer the relative productivity of a country in a product if the country actually makes the product but cannot infer the productivity if the country does not (Deardorff, 1984; Costinot et al., 2012) [1] This is an important shortcoming as there are many instances in which it would be useful to infer the productivity level that a country would enjoy in products that it does not currently make.

In addition, current Ricardian models assume that the relative productivity parameters across industries are uncorrelated (Dornbusch et al., 1977; Eaton and Kortum, 2002). This implies that the likely productivity of a country in trucks is independent of whether it currently has comparative advantage in cars or in coffee.

Imagine the following thought experiment. You have downloaded a dataset containing the exports by product for all countries of the world for the year 1995. However, due to some accident, your computer randomly erases a few entries in the matrix of exports by industry and location. How would you guess what those entries were if you had no additional information other than that contained in the surviving part of the matrix? The current Ricardian models of trade would not be useful in predicting the inter-industry variation in exports at the country level regarding the missing data, whether the industry existed or not in the real data.

In this paper we extend the Ricardian model by assuming that technological relatedness causes relative productivities to be differentially correlated across industries in a manner that can be empirically estimated. This structure implies that the comparative advantage of a country in a product can be estimated from its comparative advantage in technologically related products, even for the products the country does not currently export. This also implies that information about the relative productivities of countries with similar technological orientation should be informative of the relative productivities of industries in a given country. Hence, we can infer the similarity in the technological orientation of countries from the similarity in their output or export structure. Symmetrically, the intensity with which

[1] Deardorff (1984), quoted by Costinot et al. (2012) says that *"If relative labor requirements differ between countries, as they must for the model to explain trade at all, then at most one good will be produced in common by two countries. This in turn means that the differences in labor requirements cannot be observed, since imported goods will almost never be produced in the importing country."*.

a country exports a product should be related to the intensity with which it exports similar products, where product similarity is calculated from the pattern of co-exports of pairs of products across countries. We use these similarity measures to generate predictors of the *implied comparative advantage* of a country in an industry and show that it is strongly predictive of the *revealed comparative advantage of that country* in that industry. In addition, these estimates are strongly predictive of future changes in comparative advantage, whether among industries that already exist in a particular location or among those that have yet to emerge. In terms of our thought experiment, our approach allows us to make estimates of the missing data, and the error terms of our prediction are not just noise, but are actually predictive of future changes.

The Ricardian model can be seen as a reduced form of a more structural model that determines the productivity parameter of the labor inputs. One such model is a generalized Heckscher-Ohlin-Vanek (HOV) (Vanek, 1968) model where, implicitly, the labor productivity parameter is the consequence of the availability of an unspecified list of other factors of production. These may include many varieties of human capital, geographic factors and technological prowess, among many others. In the Appendix, we show that the essential results and reduced form equations of our approach can be derived from this setting. In an HOV interpretation, the revealed comparative advantage of a country in a product can be inferred from its revealed comparative advantage in products that have similar production functions or locations that have similar factor endowments. Interestingly, this can be derived without information regarding production functions or factor endowments.

Our results are not mainly about international trade: we obtain similar results when we use sub-national data on wage bill, employment or the number of establishments for the US, India and Chile. Clearly, a city is an economy that is open to the rest of its country and, hence, the logic behind trade models should be present, albeit with more factor mobility than is usually assumed in trade models. Our results operate both at the intensive and the extensive margins of trade: they correlate with future growth rates of country-product cells, as well as with the appearance and disappearance of new industries in each country.

Related literature

This paper is related to several strands of literature. On the one hand, it is related to the literature on the Ricardian models of trade (Dornbusch et al., 1977; Eaton and Kortum, 2002; Costinot et al., 2012), where we abandon the assumption of an absence of systematic correlations of relative productivity parameters between industries. For example, Eaton and Kortum (2002) assumed that the productivity parameters are drawn from a Frechét

distribution, except for a common national productivity parameter. Costinot et al. (2012) relaxed this assumption by assuming a country-industry parameter, but no correlation across industries in the same country. These assumptions are clearly rejected by the data, as there is very significant correlation across industries in the same country. In our results, we show that there is a systematic correlation in the patterns of comparative advantage across pairs of industries across all countries. We also show that there is a systematic correlation of the patterns of comparative advantage between pairs of countries across all industries.

We assume instead that technological relatedness across industries causes relative productivities to be correlated. The patterns we observe in the data allow us to derive implied comparative advantage estimates. It has the advantage of being able to estimate relative productivities for industries that have zero output. Moreover, the implied parameters estimated are strongly correlated with future relative productivities implying that they capture something more fundamental than the relative productivities that are calculated from contemporaneous trade.

Moreover, the previous Ricardian literature cannot infer relative productivities of industries that do not exist. An exception is Costinot and Donaldson (2012) where they estimate implied or counter-factual productivity parameters for agricultural industries using agronomic models and data. This approach requires a detailed knowledge of agricultural production functions and hence cannot be easily extended to other industries. Our approach can be extended to all industries.

This paper is also related to the controversy surrounding the Leontief Paradox. For analytical tractability, economic models are often written with few factors of production and are then extended to see if the theorems derived in the simpler setting hold for an arbitrary number of factors. But to test theories empirically, it has been necessary to take a stand on the relevant factors of production in the world. In his seminal papers, Leontief found evidence against the Heckscher-Ohlin prediction that the basket of exports of a country should be intensive in the relatively more abundant factors (Leontief, 1953, 1956). He did so by decomposing the factor content into two factors: capital and labor. Testing a multi-factor world required an extension of the Heckscher-Ohlin model, derived by Vanek (1968). The question then moved onto which factors to take into account when testing the theory empirically. [2] In most cases, it was not possible to list all factors related to the production

[2]This opened up a long literature on the relative factor content of trade (Antweiler and Trefler, 2002; Bowen et al., 1987; Conway, 2002; Davis et al., 1997; Davis and Weinstein, 2001; Deardorff, 1982; Debaere, 2003; Hakura, 2001; Helpman and Krugman, 1985; Leamer, 1980; Maskus and Nishioka, 2009; Reimer, 2006; Trefler, 1993, 1995; Trefler and Zhu, 2000, 2010; Zhu and Trefler, 2005). For example, Bowen et al. (1987) test it with 12 factors. Davis and Weinstein (2001) argue that HOV, "when modified to permit technical differences, a breakdown in factor price equalization, the existence of nontraded goods, and costs of trade, is consistent with data from ten OECD countries and a rest-of-world aggregate (p.1423). Clearly, all of these

and the tests were limited to the factors that can be measured. But these models have implications about the world that need not take a stand on what are the relevant factors of the world but can eschew that issue. The thought experiment above illustrates this idea. Products that have similar production functions should tend to be co-exported by different countries with similar intensities. Countries with similar factor endowments should tend to have similar export baskets. We can use these implications of the HOV model to estimate the missing data in our thought experiment.

This paper builds on Bahar et al. (2014), Hausmann and Klinger (2006, 2007) and Hidalgo et al. (2007) but develops a theoretical framework and explores both the extensive and the intensive margins. Our results using sub-national data relate to the urban and regional economics literature. For example, Ellison et al. (2010) try to explain patterns of industry co-agglomeration by exploring overlaps in natural advantages, labor supplies, input-output relationships and knowledge spillovers. We do not try to explain co-agglomeration but instead use it to implicitly infer similarity in the requirements of industries or the endowments of locations. Delgado et al. (2010, 2012) and Porter (2003) use US sub-national data to explain employment growth at the city-industry level, using the presence of related industry clusters. Similarly, Neffke et al. (2011) show that regions diversify into related industries, using an industry relatedness measure based on the coproduction of products within plants.

Interestingly, the measures we derive are similar to the collaborative filtering models used in the computer science literature. These models try to infer, for example, a users preference for an item on Amazon based on their purchases of similar items (Linden et al., 2003), or how they will rate news articles based on the ratings of similar users (Resnick et al., 1994). Here we derive a theoretical rationale for their logic.

This paper is structured as follows. Section 1 derives our predictors using a modified Ricardian framework. Section 2 discusses the data. Section 3 presents our results for the intensive margin. Section 4 discusses our results on the growth of industries in location. Section 5 contains our results for the extensive margin. In Section 6 we conclude with a discussion of the implications of our findings.

modifications can be construed as involving other factors, such as technological factors causing measured productivity differences, factors associated with geographic location and distance that affect transport cost, or factors that go into making nontraded goods that are used in the production of traded goods. Trefler and Zhu (2010) argue that there is a large class of different models that have the Vanek factor content prediction meaning that a test of the factor content of trade is not a test of any particular model.

1 Theoretical motivation

In this section we derive measures that capture the similarity between industries and between locations using a modified Ricardian framework. As we argued in the Introduction, a standard Ricardian model of trade that assumes that the productivity parameter of a country in an industry is a random realization from a probability distribution would not be able to explain the patterns of co-location of industries in countries or of the same industry across countries. However, if one were to make a Ricardian model compatible with these observations it would need to assume that products differ in their technological relatedness and countries tend to have similar productivities in technologically related products. With this assumption we can motivate our results in a Ricardian framework as stating that a country will export a product with an intensity that is similar to that with which countries with similar patterns of comparative advantage export that product. Also it would export that product with an intensity that is similar to that with which it exports technologically related products. In the Appendix, we derive measures that capture the similarity between industries and between locations using an approach based on Heckscher-Ohlin-Vanek (HOV) theory on factor content of production.

In our Ricardian framework, we will construct a particular relation between the technological requirements of an industry and technological endowments of a location. We will assume that the efficiency with which industry i functions in location l depends on the distance between the technological requirements of industry i and technological endowments of location l. Suppose the technological requirements of the industry i are characterized by a parameter ψ_i, which is a number on the real circle with a circumference of 1, which we denote by $\mathbb{U}$. [3] The technological endowment of location l is characterized by a parameter λ_l, also on $\mathbb{U}$. Output of industry i in location l will depend on the similarity between the requirements of the industry, ψ_i, and the endowments of the location, λ_l. More concretely,

$$\widehat{y}_{il} = A_i B_l f\left(d(\psi_i, \lambda_l)\right) \tag{1.1}$$

where d is the distance on the unit circle $\mathbb{U}$, $f : [0, 0.5] \rightarrow [0, 1]$ is a strictly decreasing function with $f(0) = 1$ and $f(0.5) = 0$ and A_i and B_l are parameters that capture the relative sizes of the location and the industry. As can be observed, output will be maximized when $\psi_i = \lambda_l$. The maximum possible distance on the circle is 0.5 and when that happens, output would be zero. We can redefine the left-hand side variable by dividing each entry

[3] We chose the unit circle to avoid boundary effects of the space. For instance, for an interval like $[0, 1]$, the boundaries, 0 and 1, will introduce break points. In reality the technological space is multi-dimensional but here we introduce a one-dimensional version to illustrate our results. Our results are not sensitive to choice of the technological space.

by the expected maximum size of the industry-location pair, A_iB_l, to calculate the relative presence of industry i in location l. This can also be interpreted as a measure of revealed comparative advantage of the location in the industry:

$$y_{il} = \frac{\widehat{y}_{il}}{A_iB_l} = f\left(d(\psi_i, \lambda_l)\right) \tag{1.2}$$

In reality, we would not be able to observe $\psi_i$and λ_l directly, but we can measure y_{il}. The basic intuition is that information about ψ_i and λ_l is contained in the presence of other industries in the same location or the presence of the same industry in other locations. For example, the difference between a locations comparative advantage in two industries, i and i', is an increasing function of the distance between the ψ_i and $\psi_{i'}$. By the same token, the difference in the share of output of the same industry across two locations l and l' would be an increasing function of the difference in the λ_l and $\lambda_{l'}$.

We can generalize this intuition by taking advantage of the information contained in the share of output of all industries in all locations. Suppose we start with a matrix Y_{il} containing the shares of industry i in location l. We can calculate the correlation matrix that contains correlations of each industry pair across all locations. We define as the product space similarity matrix $\phi_{ii'}$ between two industries i and i' as the scaled Pearson correlation between Y_i and $Y_{i'}$ across all locations:

$$\phi_{ii'} = (1 + \text{corr}\{Y_i, Y_{i'}\})/2 \tag{1.3}$$

Symmetrically, we define the country space proximity matrix $\phi_{ll'}$ between two locations l and l' as the Pearson correlation betweenY_l and $Y_{l'}$ across all industries:

$$\phi_{ll'} = (1 + \text{corr}\{Y_l, Y_{l'}\})/2 \tag{1.4}$$

If we assume that $\psi_i$and λ_l are uniformly distributed on the unit circle, and if we use a specific productivity function, $f\left(d(\psi_i, \lambda_l)\right) = 1 - 4d^2(\psi_i, \lambda_l)$, then we can derive a closed form expression for the expected value of the $\phi_{ii'}$ as a monotonic function of the distance between the ψ_is (see Appendix for the details of the calculation):

$$\phi_{ii'} = 1 - 15\left(d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'})\right)^2 \tag{1.5}$$

Similarly, our location-location proximity $\phi_{ll'}$ is a monotonic function of the distance between the endowment parameters λ_l and $\lambda_{l'}$:

$$\phi_{ll'} = 1 - 15\left(d(\lambda_l, \lambda_{l'}) - d^2(\lambda_l, \lambda_{l'})\right)^2 \tag{1.6}$$

Note that for distance $d = 0$, the expected proximity would be 1. If distance d is equal to its maximum value of 1/2 then the expected proximity would be the minimum.

We conclude that these two matrices carry information about the similarity in the requirements of pairs of industries and the endowments of pairs of locations. Thus our proximity measures use the information contained in the industry-location matrix to relate the technological requirements of an industry with the endowments of a location.

1.1 Calculating the implied comparative advantage

Equipped with our industry similarity and location similarity metrics, we can now develop a metric for the implied comparative advantage of an industry in a location. Assume that we do not know the intensity of industry i in location l. However, we do know the intensity of all other industries in this location and we know the similarities between industry i and all other industries indexed by i'. One approximation would be to look at the intensity, in that location, of other highly related industries, since these should have technological relatedness or similar factor requirements and hence similar values of Y_{il}. But how many related industries should we take into account? If we base ourselves in the single most related industry, we may have the best estimate, but we may also introduce a large error. If we average over a certain number of the most related industries and weigh our results by the degree of relatedness, we may average out some of these errors. So, following this logic, our expected value of the Y_{il} would be the weighted average of the intensity of the k nearest neighbors $Y_{il'}$ (Sarwar et al., 2001) where the weights are given by the proximity parameters $\phi_{ii'}$. We refer to this variable proxying for the implied comparative advantage as the product space density:

$$\widehat{Y}_{il}^{[I]} = \sum_{i' \in I_{ik}} \frac{\phi_{ii'}}{\sum\limits_{i" \in I_{ik}} \phi_{ii"}} Y_{i'l} \tag{1.7}$$

where I_{ik} is the k nearest neighbors of industry i:

$$I_{ik} = \{i' | Rank\left(\phi_{ii'}\right) \leq k\} \tag{1.8}$$

We can also build a similar metric using the location similarity indices. With this, the implied comparative advantage of an industry in a location would be the weighted average of the intensity of that industry in the k most related locations:

$$\widehat{Y}_{il}^{[L]} = \sum_{l' \in L_{lk}} \frac{\phi_{ll'}}{\sum\limits_{l" \in L_{lk}} \phi_{ll"}} Y_{il'} \tag{1.9}$$

with the set L_{lk} defined as:

$$L_{lk} = \{l' | Rank\left(\phi_{ll'}\right) \leq k\} \tag{1.10}$$

We refer to this variable as the country space density. We will explore the degree to which the product space and country space densities can predict the actual value of the location-industry cells using a toy model where we exactly know all the underlying parameters.

1.2 Simulating the estimators on a toy model

We illustrate how well density variables for implied comparative advantage based on the presence of related industries in the same row or the value of the same industry in related columns predict the value of each entry in the Y_{il} matrix by simulating a toy model with 100 countries and 100 products and assume a uniform distribution of the ψ_i and the λ_l on the unit circle $\mathbb{U}$. In the toy model, we exactly know the underlying parameters; hence, we can experiment with the model choice parameters. First, we verify that our industry similarity index captures the distance between the factor requirements of industries, and that our location similarity index captures the distance between the factor endowments of locations. Next, we estimate how well our density measures predict the output of each industry-location. We will then study the impact of different neighborhood filters at different levels of noise.

We first use our variables for implied comparative advantage to estimate the intensity of output of each industry-location cell. To do this, we estimate the product space density of industry i in location l by calculating the weighted average of the intensities of the k most similar industries in location l with the weights being the similarity coefficients of each industry i' to industry i. We also calculate the country space density of industry i in location l by estimating the weighted average of the intensity of industry i across the k most similar locations. Setting $k = 50$ and iterating the simulation through 5,000 trials, we find that our hybrid density model (i.e., a regression including both industry density and location density) is a powerful predictor of industry-location output (mean $R^2 = 0.784$, with 95% confidence interval of 0.715 0.853 across all simulations). However, we need not fix the neighborhood filter at $k = 50$. In Figure 1, the uppermost line shows the effect of neighborhood size on the R^2. We see that the highest R^2 value is found at $k = 4$.

This result implies that it is possible to predict the value of any entry in the Y_{il} matrix looking at the presence of related industries in the same row or the value of the same industry in related columns. This in itself is an interesting implication of our approach. But, as we will show in Section 4, not only do the product space and country space densities perform

well at predicting the Y_{il} matrix, but more surprisingly, the errors in the relationship between actual and fitted values of the Y_{il} matrix are predictive of future growth, both when looking at the intensive margin as well as the extensive margin. It is as if the rest of the matrix has more information about what the value of a cell should be than the cell itself and deviations from this expectation are corrected through subsequent growth or decline.

Finally, we can extend our simulation to examine the effect of noise in the observed output. Until now, we have assumed that the output of an industry-location, $\tilde{Y}_{il}$, is determined solely is determined solely by the distance between the technological requirement of the industry ψ_i and the technological ability of the location λ_l. We can call this the equilibrium output. Let us assume instead that the output of each industry-location can deviate from this equilibrium value because of a disturbance term ε_{il} that is normally distributed. We will explore the possibility that the disturbance term enters either linearly or exponentially. As a result of these assumptions, we no longer observe the equilibrium output Y_{il}, but instead observe only the current output, $\tilde{Y}_{il}$.

$$\tilde{Y}_{il} = Y_{il} + \varepsilon_{il} \tag{1.11}$$

Because the error term is not correlated across location or industry, we can expect that averaging our density index over several neighbors will reduce the effect of noise on our results. That is, we can achieve a better estimate of the noise-free output Y_{il} by averaging the observed, noisy output $\tilde{Y}_{il}$ of the most similar industries and locations, since the error in their output levels might cancel out. Our simulations confirm this hypothesis. We test three levels of noise in the output. Given that the standard deviation of Y_{il} in our surrogate data is 1.994 (median value from 5,000 trials) we set the standard deviation of the noise term to 1, 2 and 4 which are, respectively half, the same or twice the standard deviation of Y_{il}.

Now that the observed output incorporates an error component over the equilibrium output, the density variables are better estimates of the underlying fundamental parameters ψ_i and the λ_l than the parameters that would be inferred using the actual production. We illustrate this using a simulation of our toy model with 100 locations and 100 industries, where we now vary the standard deviation of the error term. We can then use the above formulas to calculate simulated output, proximities, and densities, setting $u = v = 50$. Figure 1 illustrates the explanatory power of our three density variables both for the additive and the exponential error models. We graph the correlation between observed output and equilibrium output as a measure of how well the model is able to implicitly capture the values of the fundamental variables ψ_i and the λ_l. When the error term has a standard deviation near zero, observed output is almost perfectly related to the underlying equilibrium output as estimated using the density variables. However, as the size of the error term increases, the

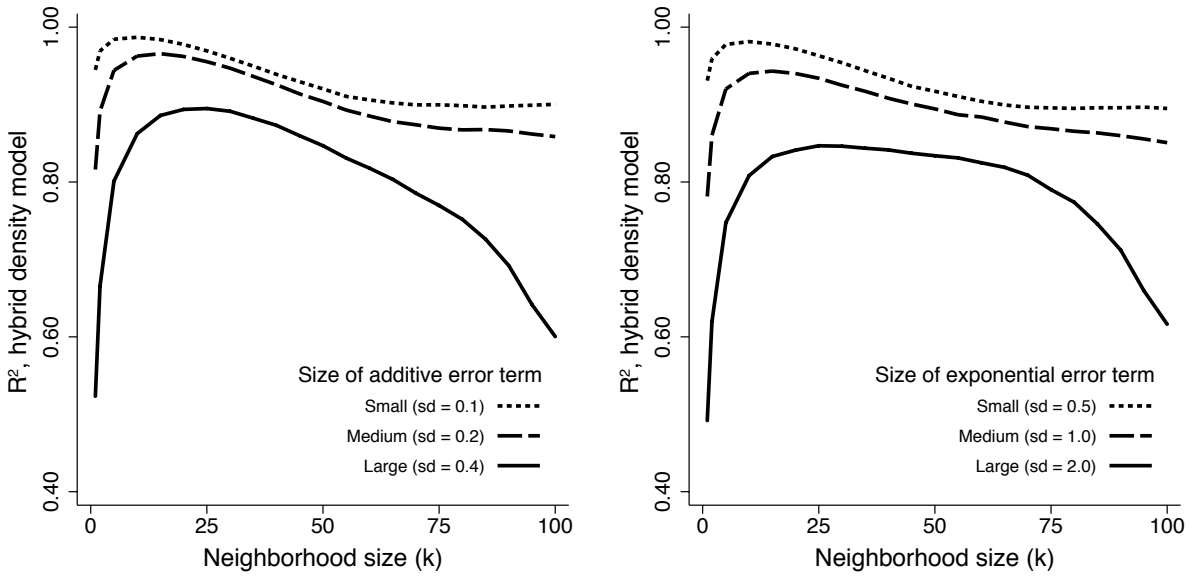

Figure 1: *Simulation of association between underlying output and hybrid density model, by size of neighborhood and noise level.*

observed output becomes increasingly less correlated with equilibrium output. The density variables are better able to capture the underlying structural variables and hence are better able to predict equilibrium output, with the Hybrid density outperforming either the product space or the country space densities because they average over a broader set of observations.

In Figure 1, we see the effect of increasing the size of the error term on the correlation between the density variables and the actual product intensity. First, we note that, as expected, a larger error term does reduce the R^2 of our estimates, though the decline is relatively small. Second, as noise increases, the R^2 peak tends to move toward mid-range k values, suggesting that the tradeoff between focusing on more related industries and averaging over a broader set of observations moves in favor of the latter. At the same time, the relationship between k and R^2 levels out as noise increases. For example, with a noise level of 2, the R^2 curve is fairly flat with predictive power roughly equal between k values of 4 and 150. When the neighborhood size gets larger, the predictive power decreases because the measure of density incorporates increasingly irrelevant information. This result suggests that finding the optimal neighborhood size may not be a first-order concern for our empirical tests.

2 Data and Methods

We now turn to the application of our approach to real data using both international and subnational datasets, which cover different countries, time periods and economic variables. After constructing our density indices, we separate our analysis between the exploration of the intensive and extensive margins. We study the growth rates of industry-location cells, which can only be defined for cells that start with a nonzero output. We study the extensive margin by looking at the appearance of industries that were not initially hosted in a particular location. For each analysis, we fit the density variables for the implied comparative advantage to current output levels, and then conduct out-of-sample regressions to explain either output growth or the appearance and disappearance of industries.

2.1 Data

We use the export dataset of countries published in *Base pour l'Analyse du Commerce International* (BACI), a database of international trade data by the *Centre d'Etudes Prospectives et d'Informations Internationales* (CEPII) (Gaulier and Zignago, 2010). Exports are disaggregated into 1,241 product categories according the Harmonized System four-digit classification (HS4), for the years 1995-2010. We restrict our sample to countries with population greater than 1.2 million and total exports of at least $1 billion in 2008. We also remove Iraq (which has severe quality issues) and Serbia-Montenegro, which split into two countries during the period studied. We drop one product, "Natural cryolite or chiolite" (HS4 code 2527), as its world trade falls to zero after the year 2006. These restrictions reduce the sample to 129 countries and 1240 products that account for 96.5% of world trade and 96.4% of the world population.

In addition to the international trade data, we test our model on three national datasets that quantify the presence of industries in locations within countries. We use the US Census County Business Patterns (CBP) database from 2003-2011. It includes data on employment and number of establishments by county, which we aggregate into 708 commuting zones (CZ; Tolbert and Sizer (1996)), and 1086 industries (NAICS 6-digit). This dataset also provides annual payroll data for 698 CZ and 941 NAICS6 industries. [4] Our Chile dataset comes from the Chilean tax authority, *Servicio de Impuestos Internos*, and includes the number of establishments based on tax residency for 334 municipalities and 681 industries, from

[4]The discrepancy between employment and establishment versus payroll sample sizes comes from the data suppression methods of Census Bureau. To protect the privacy of smaller establishments, the CBP occasionally discloses only the range of employment of an industry in a location, e.g., 1 to 20 employees. In these censored cases, we use the ranges midpoint as the employment figure (see Glaeser et al. (1992)). However, the CBP offers no payroll information in these cases, leaving a smaller payroll sample.

2005 to 2008 (Bustos et al., 2012). Lastly, we study India's economic structure using the Economic Census, containing data on employment for 371 super-districts and 209 industries, for the years 1990, 1998 and 2005. [5] For all the datasets above, we include only industries and regions that have non-zero totals for each year. This approach effectively removes discontinued or obsolete categories.

2.2 Constructing the model variables

First, we build the similarity and density indices for the implied comparative advantage introduced above for each dataset. Our first step is to normalize the export, employment and payroll data to facilitate comparison across locations, industries and time. We use the exports per capita as a share of the global average in that industry. This can be seen as a variant of Balassa's revealed comparative advantage (RCA) index (Balassa, 1964), but we use the population of a location as a measure its size rather than its total production or exports (Bustos et al., 2012) . This eliminates the impact of the movement in output or prices of one industry on the values of other industries. Specifically, we define R_{il} as:

$$R_{il,t_0} = \frac{x_{il,t_0}/pop_{l,t_0}}{\sum_l x_{il,t_0} / \sum_l pop_{l,t_0}} \tag{2.1}$$

where pop_l is the population in location l, and t_0 is the base year. Note that locations with very low populations will tend to have higher R_{il} values. To address the potential bias introduced by low-population locations, we cap R_{il} at $R_{max} = 5$, when building our similarity indices (Equations 2.2 and 2.3 below). We do not normalize the data for the number of establishments.

At this point, we can use the normalized industry intensity values, R_{il}, to build the similarity indices defined above:

$$\phi_{ii'} = (1 + \text{corr}\{R_i, R_{i'}\})/2 \tag{2.2}$$

$$\phi_{ll'} = (1 + \text{corr}\{R_l, R_{l'}\})/2 \tag{2.3}$$

In other words, two industries are similar if different locations tend to have them in similar proportions. Likewise, two locations are similar if they tend to harbor the same industries with a similar intensity. Though we use the Pearson correlation here, we obtain comparable results using other similarity measures, namely cosine distance, Euclidean distance, the Jaccard index, minimum conditional probability (Hidalgo et al., 2007) and the Ellison-Glaeser

[5] This dataset was constructed by Sam Asher and Paul Novosad and kindly shared it with us.

co-agglomeration index (Ellison and Glaeser, 1999).

Tables 1 and 2 show the top ten most similar pairs of countries and products in 2010. We note that the most similar are countries in close geographic proximity, a phenomenon that can be explained by geological and climate effects as well as regional knowledge spillovers (Bahar et al., 2012). The list of most similar pairs of products is dominated by machinery products, especially those in the "Boilers, Machinery and Nuclear Reactors," category (HS2 code 84). This matches the observation in Hausmann et al. (2011) that the machinery-related industries are highly interconnected.

Table 1: Most similar location pairs, international trade, 2010

Location l		Location l'		Location Similarity
COD	Congo, DR	COG	Congo	0.8081
CIV	Cte d'Ivoire	CMR	Cameroon	0.7987
CIV	Cte d'Ivoire	GHA	Ghana	0.7844
SWE	Sweden	FIN	Finland	0.7640
KOR	South Korea	JPN	Japan	0.7631
SDN	Sudan	ETH	Ethiopia	0.7622
KHM	Cambodia	BGD	Bangladesh	0.7543
LTU	Lithuania	LVA	Latvia	0.7526
GHA	Ghana	CMR	Cameroon	0.7519
DEU	Germany	AUT	Austria	0.7499

Table 2: Most similar industry pairs, international trade, 2010

Industry i		Industry i'		Industry Similarity
8481	Valves	8413	Liquid Pumps	0.9808
8409	Engine Parts	8483	Transmissions	0.9808
8485	Boat Propellers	8484	Gaskets	0.9784
8481	Valves	8409	Engine Parts	0.9754
7616	Aluminium Products	7326	Iron Products	0.9752
8481	Valves	8208	Cutting Blades	0.9747
8483	Transmissions	8413	Liquid Pumps	0.9747
8413	Liquid Pumps	8409	Engine Parts	0.9745
8208	Cutting Blades	8207	Interchangeable Tools	0.9743
8503	Electric Motor Parts	7326	Other Iron Products	0.9740

Having built our similarity indices, we can use them to recreate our density indices that we use to calculate implied comparative advantage with from equations 1.7 and 1.9, replacing the y_{il,t_0} with R_{il,t_0}:

$$w(u)_{il}^{[PS]} = \sum_{i' \in I_{iu}} \frac{\phi_{ii'}}{\sum_{i" \in I_{iu}} \phi_{ii"}} R_{i'l,t_0} \tag{2.4}$$

where I_{iu} is the u nearest neighbors of industry i. Similarly

$$w(v)_{il}^{[CS]} = \sum_{l' \in L_{lv}} \frac{\phi_{ll'}}{\sum_{l" \in L_{lv}} \phi_{ll"}} R_{i'l,t_0} \tag{2.5}$$

with the set L_{lv} is is the v nearest neighbors of location l. As before, we set the neighborhood sizes u and v to the 50 nearest neighbors in all cases.

3 Estimating the initial industry-location cells from the values of all other industry-location cells

As argued above, the density variables derived above are the expected value of the output intensity of any cell, given the values of other cells. To see how well they fit, we estimate the following equation:

$$\log(R_{il,t_0}) = \alpha + \beta_{PS} \log\left(w(u)_{il}^{[PS]}\right) + \beta_{CS} \log\left(w(v)_{il}^{[CS]}\right) + \varepsilon_{il,t_0} \tag{3.1}$$

where ε_{il,t_0} is the residual term.

Table 3: OLS regression of international exports by industry-location, 1995

	(1)	(2)	(3)
	Exports, 1995 (Revealed Comparative Advantage, log)		
Product Space Density (log), 1995	0.956***		0.864***
	(0.013)		(0.019)
Country Space Density (log), 1995		1.529***	0.253***
		(0.079)	(0.032)
Adjusted R^2	0.635	0.402	0.641

$N = 94,029$. Country-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3 shows that both the product space density and the country space density terms are highly significant and explain a very large fraction of the variance of the country-product

export intensity. The product space density generates a significantly higher R^2 than the country space density. Together, they explain nearly two thirds of the variation in export intensity. Table 4 shows the regressions for the US, India and Chile datasets. In all, both product space density and country space density are significant, and R^2 values are substantial. This suggests that the value of an industry-location cell in our data can be estimated with some accuracy based on the values of other industry-location cells in the matrix. However, as with any estimation, errors are made. Are these errors just noise or do they carry information about the evolution of the system? We turn to this question in the next section.

Table 4: OLS regression of initial employment, payroll and establishments by industry-location.

	(1)	(2)	(3)	(4)	(5)
	USA, 2003 employees	USA, 2003 payroll	India, 1990 employees	USA, 2003 establishments	Chile, 2005 establishments
	Revealed comparative advantage (log)			log	log
Product Space density (log)	0.620***	0.476***	0.530***	0.293***	0.306***
	(0.010)	(0.016)	(0.018)	(0.021)	(0.015)
Country Space density (log)	0.556***	0.363***	0.759***	0.517***	0.531***
	(0.010)	(0.018)	(0.013)	(0.018)	(0.011)
Observations	279,439	89,378	49,651	278,946	50,373
Adjusted R^2	0.267	0.355	0.376	0.787	0.601

Country-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4 Growth regressions

The fact that our density variables for the implied comparative advantage can explain current production is interesting, but more surprising is the fact that the residual is informative of future industry-location growth. Formally, we test for this by regressing the growth rate of the industry-location pair on the residuals that we obtain from the first stage introduced in the previous section. We construct the variables as follows. We use the standard definition of the annualized growth rate of x_{il}:

$$\dot{x}_{il} = \frac{1}{t_1 - t_0} \log \left(x_{il,t_1} / x_{il,t_0} \right) \tag{4.1}$$

where t_0 and t_1 are the initial year and final year, respectively. However, there are a large

number of locations with initial output of zero for which we cannot define a growth rate. Likewise, cases in which final output is zero (i.e., $x_{il,t_1} = 0$) are also problematic because it introduces a hard boundary that would bias the estimates. We manage these issues by separately analyzing the intensive and extensive margins. In this section, we examine growth in the intensive margin by restricting our regression sample of industry-locations to those in which $x_{il,t_0} \neq 0$ and $x_{il,t_1} \neq 0$. In Section 6, we use a *probit* regression model to examine the probability of industry appearance (i.e., growth from zero) and disappearance (i.e., collapse into zero).

Our growth regression takes the following form:

$$\dot{x}_{il} = \alpha + \beta_\varepsilon \varepsilon_{il,t_0} + \gamma c_l + \delta d_i + e_{il} \tag{4.2}$$

where a is the constant, β_ε is the regression coefficient on the residual, γ and δ are the coefficients on location and industry control variables and e_{il} is the error term of the regression.

Table 5 shows a set of growth regressions using our international export data. The dependent variable is the growth rate in the industry-location cell. The first three columns in Table 5 use as independent variable the error terms from the three regressions in Table 3. They show that the residual using both product space and country space densities, as well as both of them combined are highly significant predictors of growth and explain between 15 and 19 percent of the variance of growth between 1995 and 2010. Interestingly, in this case, the error generated by the country space density, which had lagged the product space density as an estimator of current output, does a slightly better job at predicting subsequent growth.

We now look at the robustness of these equations with respect to the inclusion of other relevant industry and location variables. First we include some basic controls regarding the initial global size of the industry in question as well as the total initial exports and population of the location. Note that these variables are all about the base year of the regression. Column 4 shows that these variables, on their own, are significantly related to subsequent growth, as noted by Glaeser et al. (1992) but as Column 5 indicates, they do not substantially affect the magnitude and significance of the density residuals.

Next, we control for the growth of the location by assuming a constant rate of growth for the location. Similarly, we introduce a control for the overall growth rate of the industry in all locations. We name these controls as radial growth variables. We show that the information captured by the error term is orthogonal to radial growth of the industry and the location. To express radial industry growth, we first calculate the global industry growth rate, $\dot{b}_i$, as the rate of growth for each industrys total (summed across all locations):

$$\dot{b}_i = \frac{1}{t_1 - t_0} \log \left(\frac{\sum_l x_{il,t_1}}{\sum_l x_{il,t_0}} \right) \tag{4.3}$$

Likewise, we calculate the average location growth rate, $\dot{b}_l$, by adding up all the industries in each location and calculating the location growth rate:

$$\dot{b}_l = \frac{1}{t_1 - t_0} \log \left(\frac{\sum_i x_{il,t_1}}{\sum_i x_{il,t_0}} \right) \tag{4.4}$$

Note that these variables would account for all the variance in growth rates of the industry-country cell if all industries within a country grew at the same rate or if all countries maintained their industry market share in the world. Deviations from balanced location growth mean that some industries are increasing or decreasing their share in the locations exports. Deviations from global industry growth mean that countries are changing their global market share in that industry. We use these balanced growth variables in two ways. First, they are an intuitive benchmark comparator for our density indices, as they represent an alternative theory of growth dynamics. Second, they are also useful to determine to what extent the density variables related to implied comparative advantage are capturing a dynamic that is orthogonal to balanced radial growth. Note, however, that the density variables are calculated with only base year data, but radial growth uses information about what happened during the 1995-2010 period.

Column 6 shows the effect of radial growth and initial size variables on subsequent growth, and as expected, they are all statistically significant and economically meaningful. Column 7 includes these variables together with the density variables. The latter substantially maintain their economic and statistical significance while they increase the R^2 relative to column 6 by about 26% (from 0.241 to 0.305).

Table 5: OLS regression of employment, payroll and establishments growth by industry-location

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Growth in exports (log), 1995-2010								
Residual, Product Space density, 1995	-0.029*** (0.001)		-0.014*** (0.002)		-0.013*** (0.002)		-0.022*** (0.002)		-0.027*** (0.001)
Residual, Country Space density, 1995		-0.024*** (0.001)	-0.016*** (0.002)		-0.009*** (0.002)		-0.005*** (0.001)		-0.009*** (0.001)
Industry-location exports (log), 1995				-0.018*** (0.001)	-0.008*** (0.001)	-0.016*** (0.001)	-0.003*** (0.001)	-0.027*** (0.001)	0.002 (0.001)
Location population (log), 1995				0.017*** (0.003)	0.010*** (0.003)	0.016*** (0.002)	0.006** (0.002)		
Global industry total (log), 1995				0.021*** (0.001)	0.011*** (0.002)	0.016*** (0.001)	0.005*** (0.001)		
Radial industry growth (log), 1995-10						0.373*** (0.008)	0.378*** (0.007)		
Radial location growth (log), 1995-10						0.156*** (0.036)	0.269*** (0.036)		
Industry FE								Yes	Yes
Location FE								Yes	Yes
Adjusted R^2	0.156	0.178	0.196	0.169	0.211	0.241	0.305	0.381	0.407

$N = 94,029$. Country-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

In addition to the balanced growth variables, we also test our model using industry and location fixed effects. These capture all industry and location effects, subsuming the size and balanced growth variables as well as any other source of variation at the location or the industry level. Thus, any additional explanatory power after controlling for the initial size of the industry-country cell and these fixed effects must come entirely from industry-location interactions.

Column 8 shows a growth equation with both country and industry fixed effects as well as the initial country-industry size. Column 9 reintroduces the density variables and shows that their economic and statistical significance in undiminished. It is important to point out that the product and country space residuals can be calculated using only base year data and no information regarding current growth, while the coefficients on the fixed effects can only be calculated ex post. This means that the residual of the first stage regression of the density variables are strongly predictive of non-radial growth in the subsequent 15 years.

The robust and negative signs in the Columns 4-8 for initial industry-location exports confirm Rodrik (2013)'s observation of unconditional convergence at the industry level. But the significance of our density measures imply a richer structure in the convergence patterns of countries.

Next, we apply the same process to our US, Chile and India datasets, over the maximum period for which we have data (Table 6). We find that the product space and country space residuals are highly significant predictors of industry-location growth, both before and after controlling for initial output, industry and location size, and balanced growth ($p < 0.01$ for all cases).

Lastly, we find that our results are robust to the choice of base year and end year (t_0 and t_1, respectively). Figure 2 shows the adjusted R^2 values for our international trade regressions over all possible year combinations. Each regression explains a sizable portion of the variation in export growth, with the lowest adjusted R^2 exceeding 7.6%, and a mean R^2 of 14%. Interestingly, we find that predictive power improves as the prediction interval increases. This indicates that the density indices do not capture a short-term mean reversion effect, but a longer-term phenomenon.

4.1 Double-out-of sample robustness check

In order to calculate our density variables for the implied comparative advantage we use information that involve the values of cells other than the one to be estimated. However, other information regarding that location or that industry is also used in the calculations. This may create some concerns regarding endogeneity. We can address this issue by splitting

our data into a training set and a testing set, a process referred to as cross validation in the computer science literature. We only use information from the training sets to build our density indices. For the product space, we estimate the similarity between industries using half of the locations. Likewise, for the country space, we estimate the similarity between locations using half of the industries. This approach leaves one quarter of the industry-location observations completely outside of the sets we used to build our similarity indices. Finally, we use these similarity indices to build density indices for the testing set. Having built our out-of-sample predictors, we can repeat the regressions using only on the testing data.

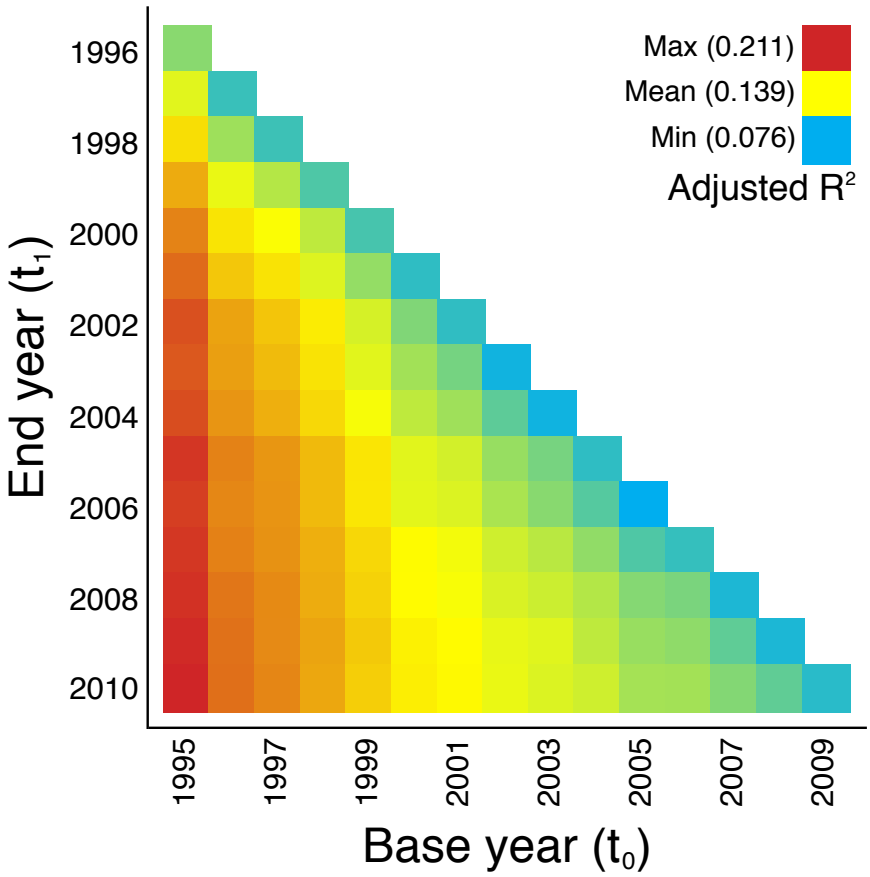

Figure 2: *Heat map of out of sample predictions of export growth, hybrid density model.*

Table 6: OLS regression of employment, payroll and establishments growth by industry-location.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	USA, 2003-2011						Chile, 2005-2008		India, 1990-2005	
	Employment growth		Establishments growth		Payroll growth		Establishments growth		Employment growth	
Residual, Product Space density	-0.018*** (0.001)	-0.027*** (0.001)	-0.007*** (0.001)	-0.002*** (0.001)	-0.012*** (0.001)	-0.045*** (0.001)	-0.027*** (0.001)	-0.031*** (0.005)	-0.209*** (0.008)	-0.264*** (0.009)
Residual, Country Space density	-0.026*** (0.001)	-0.025*** (0.001)	-0.022*** (0.001)	-0.024*** (0.001)	-0.034*** (0.002)	-0.024*** (0.001)	-0.009*** (0.002)	-0.020*** (0.002)	-0.104*** (0.009)	-0.064*** (0.007)
Initial industry -location level (log)		0.008*** (0.001)		-0.003*** (0.000)		0.016*** (0.001)		0.007 (0.005)		-0.013 (0.008)
Initial location population (log)		-0.014*** (0.001)		0.003*** (0.000)		-0.026*** (0.002)		0.014*** (0.002)		-0.017 (0.012)
Initial global industry total (log)		-0.009*** (0.001)		-0.000 (0.000)		-0.023*** (0.002)		0.023*** (0.001)		0.039*** (0.009)
Radial industry growth (log)		0.383*** (0.004)		0.331*** (0.004)		0.379*** (0.004)		0.681*** (0.013)		1.047*** (0.010)
Radial location growth (log)		0.324*** (0.013)		0.283*** (0.009)		0.210*** (0.012)		0.423*** (0.042)		0.519*** (0.079)
Observations	279,439		278,946		89,378		50,373		49,651	
Adjusted R^2	0.150	0.213	0.092	0.226	0.162	0.340	0.059	0.310	0.199	0.427

Location-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Tables 7 and 8 apply this process to our international trade dataset over the full 1995-2010 period. We find that the explanatory power of our out-of-sample hybrid model is comparable to that of the in-sample model (R^2 values are 62.2% and 55.7% for regressions of current export levels, and 18.7% versus 18.5% for regressions of export growth). Furthermore, adding the in-sample density terms to the out-of-sample dataset yields a negligible marginal contribution to R^2. Finally, combining the in-sample and out-of-sample predictors shows a marginally higher R^2 but with drastically reduced significance, indicating a high degree of co-linearity between the two types of variables. This suggests that endogeneity is not driving our results.

Table 7: Out-of-sample OLS regression of international exports by industry-location, 1995.

	(1)	(2)	(3)
	Exports, 1995 (revealed comparative advantage, log)		
Product Space density (log) out-of-sample, 1995	0.916*** (0.025)		0.940*** (0.065)
Country Space density (log) out-of-sample, 1995	0.150*** (0.038)		0.063 (0.046)
Product Space density (log) in-sample, 1995		0.830*** (0.029)	-0.035 (0.066)
Country Space density (log) in-sample, 1995		0.357*** (0.049)	0.121** (0.053)
Adjusted R^2	0.622	0.557	0.622

$N = 23,794$. Country-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

5 The extensive margin: Discrete industry appearances and disappearances

In previous sections we analyzed the rate of growth of exports, employment, payroll and number of establishments in industry-locations that already exist. In this section, we focus on the extensive margin, looking at the *appearance* and *disappearance* of industries in locations.

To do this, we first need to establish which industry-locations are present and which are absent. The case is simple when using the US and Chilean datasets because they report the

number of establishments. In these cases, an industry is present in a location if at least one establishment is reported to exist there. Formally, we capture this signal with the binary presence variable M_{il}:

$$M_{il,t_0} = \begin{cases} 1 & x_{il,t_0} \geq 1 \\ 0 & x_{il,t_0} = 0 \end{cases} \tag{5.1}$$

where, as before, x_{il,t_0} is the number of establishments in industry i and location l in year t_0. In this notation, we refer to an industry location as present when $M_{il,t_0} = 1$ and absent when $M_{il,t_0} = 0$. Likewise, an appearance between years t_0 and t_1 is defined as $M_{il,t_0} = 0 \rightarrow M_{il,t_1} = 1$, while a disappearance is defined as $M_{il,t_0} = 1 \rightarrow M_{il,t_1} = 0$.

Table 8: Out-of-sample OLS regression of growth in international exports by industry-location, 1995-2010

	(1)	(2)	(3)
	Growth in exports (log), 1995-2010		
Residual, Product Space density, out-of-sample, 1995	-0.012*** (0.002)		-0.006*** (0.002)
Residual, Country Space density, out-of-sample, 1995	-0.014*** (0.002)		-0.009** (0.004)
Residual, Product Space density, in-sample, 1995		-0.012*** (0.002)	-0.006*** (0.002)
Residual, Country Space density, in-sample, 1995		-0.014*** (0.002)	-0.006 (0.003)
Adjusted R^2	0.187	0.185	0.189

$N = 23,794$. Country-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

To study the extensive margin in the international trade dataset we need to decide on an equivalent definition of presence and absence. Here, the concern is that the data may include errors that imply the presence of an industry when it is simply a case of small re-exports or clerical error. We define an industry to be absent in a location if $R_{il,t_0} < 0.05$, meaning that exports are less than 1/20th of the average per capita exports for the world. We will consider an industry to be present if R_{il} is above 0.25. We will define an appearance as a move from $R_{il,t_0} < 0.05$ to $R_{il,t_1} > 0.25$ and a disappearance as a move from $R_{il,t_0} > 0.25$ to $R_{il,t_0} < 0.05$ as originally used by Bustos et al. (2012).

Table 9: Probit regression of industry-location extensive margin, US, Chile and International

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	USA (establishments) Industry presences in 2003			Chile (establishments) Industry presences in 2005			International (exports) Industry presences in 1995		
Product Space density, initial year	0.266*** (0.001)		0.022*** (0.002)	1.191*** (0.005)		1.165*** (0.006)	0.397*** (0.006)		0.306*** (0.007)
Country Space density, initial year		0.795*** (0.004)	0.772*** (0.005)		0.939*** (0.005)	0.822*** (0.006)		0.348*** (0.004)	0.160*** (0.004)
All industry-locations		768,888			227,454			159,960	
Present industries		324,622			55,347			47,337	
Presence rate		42.22%			24.33%			29.59%	
Area Under the Curve	0.924	0.940	0.940	0.815	0.900	0.911	0.933	0.859	0.914
Pseudo R^2	0.341	0.493	0.495	0.357	0.193	0.454	0.353	0.226	0.376

Location-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Thus, our definition of extensive margin change represents a fivefold increase or decrease in output around very low levels. While these thresholds are somewhat arbitrary, we obtain similar results using different thresholds.

We apply these definitions to the US and Chilean establishment data and to the international trade data. In the US, we classify 324,622 industry locations as present in 2003, or 42% of the total sample of industry locations. Of these present industries, 45,108 became absent by 2011, yielding a disappearance rate of 14%. Likewise, 37,681 industries that were absent in 2003 became present by 2011, resulting in an appearance rate of 8.5%. In Chile, 55,347 industries were present in 2005, or 24% of the sample. By 2008, 4,762 of these industries became absent (a disappearance rate of 8.6%) while 11,496 initially absent industries became present (an appearance rate of 6.7%). Internationally, 47,337 industries were present in our base year of 1995, or 29.6% of the sample. By 2010, 7,089 of these present industries became absent (a disappearance rate of 7.5%) while 3,648 initially absent industries became present (an appearance rate of 7.7%).

We can now use our density indices for the implied comparative advantage to explain the appearance and disappearance of industries by location. First, we explore use our density variables to generate an expected presence or absence estimation for each industry-location cell by using a probit model. In particular, we regress M_{il} on product space and country space density. Our probit model estimates the probability of industry presence in a location in the base year:

$$P\left(M_{il,t_0}=1\right)=\Phi\left(\alpha+\beta_{PS}w_{il}^{[PS]}+\beta_{CS}w_{il}^{[CS]}\right) \tag{5.2}$$

where Φ is a normal cumulative distribution function. Note that as for the intensive margin, the model in Equation 5.2 uses only information from the base year. Going forward, we denote the expected presence or absence of an industry in a location at time t_0 as M_{il,t_0}:

$$M_{il,t_0}=\widehat{M}_{il,t_0}+\varepsilon_{il,t_0} \tag{5.3}$$

where $\widehat{M}_{il,t_0}$ is the expected probability of industry presence and ε_{il,t_0} is the residual error term. We then use the residual to predict changes to M_{il,t_0}, i.e., industry appearances and disappearances. Our predictive criterion is that M_{il,t_0} will approach $\widehat{M}_{il,t_0}$ as time passes, that is, M_{il,t_0} approaches the values that are signaled by the country space and product space densities.

In addition to the pseudo-R^2 statistic, we evaluate the accuracy of these predictions using the *area under the receiver-operating characteristic (ROC) curve*. The ROC curve plots the rate of true positives of a continuous prediction criterion (the residual ε_{il,t_0} in our case) as a

function of the rate of false positives. The area under the curve (AUC) statistic is equivalent to the Mann-Whitney statistic (the probability of ranking a true positive ahead of a false positive in a prediction criterion). By definition, a random prediction will find true positives and false positives at the same rate, and hence will result in an $AUC = 0.5$. A perfect prediction, on the other hand, will find all true positives before giving any false positive, resulting in an $AUC = 1$.

Table 9 applies our probit regression model to the US and Chilean establishment data and international export data to the first year for which we have information in the respective datasets. In the initial regression, we see that our product space and country space density terms explain between one third and one half of the variance in industry-location. Also, coefficients on all terms are positive and highly significant, meaning that a high value for density is strongly indicative of the presence of an industry in a location. The AUC are very high (AUC between 91% and 94% for hybrid models).

Next, we use the residual term from these regressions to predict industry appearances and disappearances over the maximum period covered in each dataset (Table 10). For all cases, the coefficients are highly significant, and have the expected sign. In the US, over an 8-year period, the hybrid model predicts industry appearances with an AUC of 83% and disappearances with an AUC of 86%. For the Chilean data over a 3-year horizon, the hybrid models AUC is 80% for appearances and 72% for disappearances. For the international trade data over a 15-year horizon, the AUC is 72.3% for appearances and 74.2% for disappearances. This suggests that the "unexpectedly absent" industries tend to preferentially appear over time while the "unexpectedly present" industries tend to disappear.

6 Conclusions

In this paper we have shown that the intensity of an industry-location cell follows a pattern that can be discerned from the presence of related industries in that location (product-space density) or of that industry in related locations (country-space density). Moreover, the error term in the predicted pattern is not pure noise but instead carries information regarding the future level, and hence the growth rate, of that industry-location cell. These dynamics include components that are orthogonal to pure industry or location effects, but instead capture industry-location interactions. We have shown these results using international trade data as well as sub-national data for the USA, India and Chile. We have shown that they operate both at the intensive as well as the extensive margin, that they are not due to endogeneity in the information and that they operate at long horizons of over a decade.

Table 10: Probit regression of changes in industry-location extensive margin, US, Chile and international

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	USA (establishments) Industry appearances, 2003-11			Chile (establishments) Industry appearances, 2005-08			International (exports) Industry appearances, 1995-10		
Residual, Product Space density	-2.858*** (0.026)			-2.636*** (0.037)			-1.903*** (0.059)		
Residual, Country Space density		-3.004*** (0.017)			-1.757*** (0.038)			-1.327*** (0.032)	
Residual, hybrid density			-2.994*** (0.017)			-2.389*** (0.031)			-1.786*** (0.044)
Initially absent		444,266			172,107			94,547	
Industry appearances		37,681			11,496			7,089	
Appearance rate		8.48%			6.68%			7.50%	
Area under the curve	0.801	0.832	0.834	0.757	0.747	0.803	0.750	0.692	0.723
Pseudo R^2	0.059	0.145	0.144	0.064	0.021	0.073	0.019	0.027	0.028
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	USA (establishments) Industry disappearances, 03-11			Chile (establishments) Industry disappearances, 05-08			International (exports) Industry disappearances, 95-10		
Residual, Product Space density	2.953*** (0.018)			0.929*** (0.039)			1.213*** (0.032)		
Residual, Country Space density		2.265*** (0.009)			1.435*** (0.038)			1.630*** (0.050)	
Residual, hybrid density			2.272*** (0.009)			1.368*** (0.030)			1.265*** (0.031)
Initially present		324,622			55,347			47,337	
Industry disappearances		45,108			4,762			3,648	
Disappearance rate		13.90%			8.60%			7.71%	
Area under the curve	0.840	0.854	0.855	0.625	0.708	0.722	0.746	0.716	0.742
Pseudo R^2	0.231	0.249	0.250	0.022	0.066	0.081	0.080	0.068	0.087

Location-clustered robust standard errors in parentheses.
Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

We motivated our approach with a modified Ricardian model in which the industry productivity parameters of each location are correlated among technologically similar industries. This means that whatever determines the comparative advantage of a location in an industry also affects technologically related industries. In this case, product space density is informative of the advantage of a location in technologically related industries while country space density is informative of the presence of the industry in technologically similar locations. We can use these density variables to estimate an implied comparative advantage value. This information can be obtained even if the location does not currently hosts that industry.

Ricardian models are reduced form models, where other elements are subsumed in the labor productivity parameters. We show that we can motivate our approach also with a Heckscher-Ohlin-Vanek (HOV) model with an indeterminate number of factors of production. From an HOV point of view, the intensity of output in an industry-location cell should be related to the adequacy of the match between the factor requirements of the industry and the factor endowments of the location. Industries with similar factor requirements should be similarly present across locations while similarly endowed locations should host a similar suite of industries. Hence, the correlation between the intensity of presence of pairs of industries across all locations is informative of the similarity of their factor requirements while the correlation between output intensity of pairs of locations across all industries is informative of the similarity in their factor endowments.

An important question is why is it that our two density variables would carry information about the future level of the industry, even after controlling for location and industry effects and the overall growth rate of the location and the industry in question. One interpretation is that each industry-location cell is affected by a zero-mean independent and identically distributed (i.i.d.) shock that causes a deviation of its output from their equilibrium levels. In this interpretation, since over time the expected value of the i.i.d. shock is zero, then the underlying fundamentals become expressed and it is these that are captured by our approach. An alternative interpretation is that what we are observing is the consequence of inter-industry spillovers such as Marshallian and/or Jacobs externalities (Ellison et al., 2010; Glaeser et al., 1992). In this case, the productivity of an industry-location cell is affected by the presence of related industries through spillovers. The fact that these take time is what would explain why our predictive power peaks at time periods of a decade or more. Future research would need to test for these alternative hypotheses.

Our results are also informative for models of unbalanced growth. Much of growth theory has been based on the exploration of solutions around a balanced growth path, but there has been a growing literature that tries to cope with structural transformation, along the Kuznets facts (1973), i.e. the secular decline of agriculture in employment and output, the

rising share of services and the inverted U shaped path of manufacturing. To cope with these features, some models use non-homothetic demand, as in a minimum level of food consumption or a hierarchy of needs. [6] Other models use differential capital intensities across industries that are then rebalanced as capital deepens (Baumol (1967), Acemoglu and Guerrieri (2008). However, the stylized facts uncovered in this paper show a more subtle and fine-grained structure of predictable transformations. First, the structure is observable in exports and not just in employment and output, meaning that what drives these regularities is changes in supply rather than changes in domestic demand. Secondly, the patterns we observe are too intricate to be determined by differences in capital intensity. So this paper suggests that at least some drivers of differential growth lie elsewhere.

From a Ricardian viewpoint, the conjecture would be that mastery of specific technologies affects the productivity of related industries, a feature that is not incorporated into current Ricardian models that productivity draws are completely random (Eaton and Kortum, 2002), or sector specific (Costinot et al., 2012). Efforts to improve on one industry's productivity spillover into other related industries. The unexploited aspects of technological relatedness are reflected in the difference between a country's output structure and the international norm. These differences get diminished over time as firms exploit technological spillovers.

From an HOV perspective, the explanation requires an understanding of forces that affect the differential accumulation of multiple factors and not just their reallocation, which should happen at shorter time horizons. One conjecture is that the world is characterized by many factors of production that enter differentially in different industries with a complex set of complementarities. At any point in time, the endowment of the many factors is not consistent with an equalization of their rates of return, causing differential factor accumulation. Furthermore, given complementarity, the accumulation of one factor, in response to an initial disequilibrium will affect the return to other factors triggering further factor accumulation. The pattern of factor proportions that equalize returns is better reflected in the international average than in the country's own history. As a consequence, the output composition derived from the experience of others can be informative of the long-term trends in a particular country. Future research should test whether any of these conjectures make sense.

[6] The ample literature on non-homothetic preferences is reviewed in Matsuyama (2005).

7 Appendix

7.1 Calculation of Expected Similarity Coefficient

In this Technical Appendix, we will derive the expected similarity coefficient between two locations (products) given that the revealed comparative advantage of industry i in location l is:

$$y_{il} = 1 - 4d^2(\psi_i, \lambda_l) \tag{7.1}$$

where d is the shortest distance between independent and uniformly distributed ψ_i and λ_l parameters on a circle of perimeter 1. We can define the similarity $\phi_{ii'}$ between two industries i and i' given by

$$\phi_{ii'} = (1 + \text{corr}\{Y_i, Y_{i'}\})/2 \tag{7.2}$$

where corr is defined as

$$\text{corr}\{Y_i, Y_{i'}\} = \frac{\sum_l (y_{il} - \overline{y}_i)(y_{i'l} - \overline{y}_{i'})}{\sqrt{\sum_l (y_{il} - \overline{y}_i)^2 \sum_l (y_{i'l} - \overline{y}_{i'})^2}} \tag{7.3}$$

Since each ψ_i and λ_l are independently distributed, using law of large numbers, the sums in the correlation expressions can be converted to expectation values, namely:

$$\text{corr}\{Y_i, Y_{i'}\} = \frac{E[(y_{il} - \overline{y}_i)(y_{i'l} - \overline{y}_{i'})|\psi_i, \psi_{i'}]}{\sqrt{E[(y_{il} - \overline{y}_i)^2|\psi_i] E[(y_{i'l} - \overline{y}_{i'})^2|\psi_{i'}]}} \tag{7.4}$$

Since $\psi_i$and $\psi_{i'}$ are identical independently variables, the correlation becomes:

$$\text{corr}\{Y_i, Y_{i'}\} = \frac{E[(y_{il} - \overline{y}_i)(y_{i'l} - \overline{y}_{i'})|\psi_i, \psi_{i'}]}{E[(y_{il} - \overline{y}_i)^2|\psi_i]} \tag{7.5}$$

To make the calculations more tractable, if we use $\tilde{y}_{il} = (1 - y_{il})/4 = d^2(\psi_i, \lambda_l)$ instead of y_{il}, the similarity measure will remain the same. Using the identity:

$$E[(\tilde{y}_{il} - \overline{\tilde{y}}_i)^2|\psi_i] = E[\tilde{y}_{il}^2|\psi_i] - E^2[\tilde{y}_{il}|\psi_i] \tag{7.6}$$

we can calculate the denominator in Equation 7.5 using these separate terms. First,

$$E[\tilde{y}_{il}|\psi_i] = \int_0^1 d^2(\psi_i, \lambda_l) d\lambda_l = 2\int_0^{1/2} y^2 dy = 2[y^3/3]_0^{1/2} = 1/12 \tag{7.7}$$

and

$$E[\tilde{y}_{il}^2|\psi_i] = \int_0^1 d^4(\psi_i, \lambda_l)d\lambda_l = 2\int_0^{1/2} y^4 dy = 2[y^5/5]_0^{1/2} = 1/80 \tag{7.8}$$

hence, the denominator in Equation 7.5 becomes:

$$E[(\tilde{y}_{il} - \bar{\tilde{y}}_i)^2|\psi_i] = \frac{1}{80} - \left(\frac{1}{12}\right)^2 = \frac{1}{180} \tag{7.9}$$

We can write the numerator in Equation 7.5 as:

$$\begin{aligned} E[(y_{il} - \overline{y}_i)(y_{i'l} - \overline{y}_{i'})|\psi_i, \psi_{i'}] &= \int_0^1 \left(d^2(\psi_i, \lambda_l) - \frac{1}{12}\right)\left(d^2(\psi_{i'}, \lambda_l) - \frac{1}{12}\right) d\lambda_l \\ &= \int_0^1 [d(\psi_i, \lambda_l)d(\psi_{i'}, \lambda_l)]^2 d\lambda_l - \frac{1}{144} \end{aligned} \tag{7.10}$$

To calculate the integral, we will measure all the distances on the circle relative to ψ_i. Let's define $\Delta_{ii'} \equiv d(\psi_i, \psi_{i'})$. We can write the integral in Equation 7.10 as

$$\begin{aligned} \int_0^1 [d(\psi_i, \lambda_l)d(\psi_{i'}, \lambda_l)]^2 d\lambda_l = &\int_0^{1/2} [y(y - \Delta_{ii'})]^2 dy \\ &+ \int_{1/2}^{1/2+\Delta_{ii'}} [(1-y)(y - \Delta_{ii'})]^2 dy \\ &+ \int_{1/2+\Delta_{ii'}}^1 [(1-y)(1-y+\Delta_{ii'})]^2 dy \end{aligned} \tag{7.11}$$

The first integral in Equation 7.12 is:

$$\int_0^{1/2} [y(y - \Delta_{ii'})]^2 dy = \frac{20\Delta_{ii'}^2 - 15\Delta_{ii'} + 3}{480}$$

The second integral in Equation 7.12 is:

$$\int_{1/2}^{1/2+\Delta_{ii'}} [(1-y)(y - \Delta_{ii'})]^2 dy = \frac{16\Delta_{ii'}^5 - 80\Delta_{ii'}^4 + 160\Delta_{ii'}^3 - 120\Delta_{ii'}^2 + 30\Delta_{ii'}}{480}$$

Finally, the third integral in Equation 7.12 is:

$$\int_{1/2+\Delta_{ii'}}^{1} [(1-y)(1-y+\Delta_{ii'})]^2 dy = \frac{-16\Delta_{ii'}^5 + 20\Delta_{ii'}^2 - 15\Delta_{ii'} + 3}{480}$$

Hence

$$\int_0^1 [d(\psi_i, \lambda_l) d(\psi_{i'}, \lambda_l)]^2 d\lambda_l = \frac{-80\Delta_{ii'}^4 + 160\Delta_{ii'}^3 - 80\Delta_{ii'}^2 + 6}{480} = \frac{1}{180} - \frac{1}{6}\left(\Delta_{ii'} - \Delta_{ii'}^2\right)^2 \quad (7.12)$$

Plugging back calculated numerator and denominator into Equation 7.5, we obtain:

$$\begin{aligned} \text{corr}\{Y_i, Y_{i'}\} &= \frac{E[(y_{il} - \overline{y}_i)(y_{i'l} - \overline{y}_{i'})|\psi_i, \psi_{i'}]}{E[(y_{il} - \overline{y}_i)^2|\psi_i]} = \frac{1/180 - \left(\Delta_{ii'} - \Delta_{ii'}^2\right)^2/6}{1/180} \\ &= 1 - 30\left(\Delta_{ii'} - \Delta_{ii'}^2\right)^2 = 1 - 30\left(d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'})\right)^2 \end{aligned} \quad (7.13)$$

Then the similarity between industries i and i' is:

$$\phi_{ii'} = (1 + \text{corr}\{Y_i, Y_{i'}\})/2 = 1 - 15\left(d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'})\right)^2 \quad (7.14)$$

7.2 Motivation Based on Factor Content of Production

In the HOV tradition, the factor endowments of a location determine which industries will be present there. To set up this model, we will make following standard HOV assumptions:

1. There is full employment of all factors in each location.
2. Factor prices are equalized across all locations.
3. All locations have access to the same technologies for all industries.
4. Production technologies exhibit constant returns to scale. Note that requirements 2-4 imply that there would be a fixed optimal combination of factor inputs to produce each output.

With these assumptions, we can write the full employment condition for all factors in all locations as a linear function:

$$AY = F \tag{7.15}$$

where

- $A = N_f \times N_i$ is a matrix of factor inputs required to produce one unit of output in each industry.
- $Y = N_i \times N_l$ is a matrix where $Y_{i,l}$ represents location ls output in industry i.
- $F = N_f \times N_l$ is a matrix where $F_{f,l}$ represents location ls endowments of factor f.

From an empirical point of view, we can only observe Y the matrix of industry-location outputs. Empirically, we do not observe either the factor requirements of each industry A or factor endowments of each location F. In fact, we do not even have an exhaustive list of all factors. Following Equation 2.4 of Feenstra (2003), it is convenient to put the observable Y matrix on the left and leave the unobservable matrices on the right. In order to achieve this, we assume that $N_i = N_f$ and the A matrix is invertible. We define $B = A^{-1}$ such that $B \times A = I_{N_f}$, where I_{N_f} is the $N_f \times N_f$ identity matrix. The B matrix indicates how much output is generated by the employment of each factor in an industry. If we multiply both sides of Equation 7.15 by the B matrix, we obtain:

$$Y = BF \tag{7.16}$$

What can be inferred about the B and F matrices given that we can only observe matrix Y? Obviously, we will not be able to get information about individual elements of these matrices. Yet, we will show that the similarities in the factor requirements of two industries or the similarity between the factor endowments of two locations can be obtained from the information in the Y matrix. In subsections below, we first develop similarity measures between the factor requirements of pairs of industries and between the factor endowments of pairs of locations. This will prove instrumental for our purposes.

7.2.1 Similarities between the factor requirements of two industries

We will now derive a measure of input similarity of two industries, using Equation 7.16. We will assume that two industries, i and i', are similar if their associated row vectors in the B matrix, namely B_i and $B_{i'}$, are similar. Each element of the Y matrix can be written as:

$$Y_{il} = \sum_f B_{if} F_{fl} \tag{7.17}$$

If we denote Y_i and B_i as the row vectors of Y and B matrices, this equation can be rewritten in vector notation for all locations as:

$$Y_i = B_i F \tag{7.18}$$

We will now calculate the covariance across all locations of a given industry. For this we first need to calculate the average production of each industry. Given Equation 7.18, average production of industry i can be calculated as:

$$\overline{Y}_i = \frac{\sum_l Y_{il}}{N_l} = \sum_f B_{if} \frac{\sum_l F_{fl}}{N_l} = \sum_f B_{if} \overline{F}_f \tag{7.19}$$

where $\overline{F}_f$ is the average presence of factor f across all locations. Subtracting the last two expressions from one another, we arrive at:

$$Y_i - \overline{Y}_i = B_i(F - \overline{F}) \tag{7.20}$$

where $\overline{F}$ is a $N_f \times N_l$ matrix that repeats in each row f the average endowment of the world in that factor $\overline{F}_f$. Using Equation 7.20, we can relate the observed covariance of the rows of the Y matrix to those of the unobserved B matrix:

$$(Y_i - \overline{Y}_i)(Y_{i'} - \overline{Y}_{i'})^t = B_i(F - \overline{F})(F - \overline{F})^t B_{i'}^t \tag{7.21}$$

$C \equiv (F - \overline{F})(F - \overline{F})^t$ matrix is the covariance matrix of rows of F matrix and, by definition, it is a square and symmetric matrix. The C matrix can be written as:

$$C = U \Sigma U^t \tag{7.22}$$

where U is a unitary matrix formed by the eigenvectors of C and Σ is a diagonal matrix whose elements are eigenvalues of C. If we define $\tilde{B}_i = B_i U$, then we can write the right hand side of Equation 7.23 as

$$(Y_i - \overline{Y}_i)(Y_{i'} - \overline{Y}_{i'})^t = \sum_f \tilde{B}_{if} \tilde{B}_{i'f}^t \sigma_f \tag{7.23}$$

where σ_f is the f^{th} (largest) eigenvalue of the covariance matrix, C. In one extreme, we can assume $\sigma_f = \sigma$ for all f. This would happen, for instance, If all rows of the F matrix are independently and identically distributed (i.i.d.). An interpretation of this assumption is that locations accumulate factors separately and independently. This assumption is unlikely to be true about the world but it simplifies our proof considerably; we give evidence of the

generality of this approach in our simulations. Using this assumption, the right hand side becomes

$$\sum_f \tilde{B}_{if}\tilde{B}^t_{i'f}\sigma_f = \sigma\tilde{B}_i\tilde{B}^t_{i'} = \sigma B_i U U^t B^t_{i'} = \sigma B_i B^t_{i'} \tag{7.24}$$

Dividing both sides of Equation 7.24 by the standard deviation of Y_i and $Y_{i'}$, we can relate the correlation of the rows of the Y matrix to elements of the B matrix:

$$\text{corr}\{Y_i, Y_{i'}\} = \frac{(Y_i - \overline{Y}_i)(Y_{i'} - \overline{Y}_{i'})^t}{\sigma_{Y_i}\sigma_{Y_{i'}}} \approx \frac{\sigma}{\sigma_{Y_i}\sigma_{Y_{i'}}} B_i B^t_{i'} \tag{7.25}$$

where corr represents the Pearson correlation between vectors. Since this is a variable with a range $(-1, 1)$ we renormalize it to build a similarity metric between 0 and 1. Hence, we can estimate a measure of the similarity between the factor requirements of two industries, i and i':

$$\phi_{ii'} = (1 + \text{corr}\{Y_i, Y_{i'}\})/2 \tag{7.26}$$

Following Hausmann and Klinger (2006) and Hidalgo et al. (2007), we refer to this industry-industry similarity matrix as the product space.

7.2.2 Similarities between factor endowments of two locations

To quantify the similarities between the factor endowments of two locations, we will use an analogous approach. For two locations l and l', we would like to measure the similarity between their factor endowment vectors, F_l and $F_{l'}$. If we denote Y_l and F_l as the l^{th} column vectors of Y and F matrices respectively, the output of a location is related to its factor endowments by:

$$Y_l = BF_l \tag{7.27}$$

Note that our calculations in Section 2.1.1 can be replicated here because if we take the transposes of both sides in Equation 7.27, we will arrive to an expression similar to Equation 7.18. Assuming that the columns of B matrix are independently and identically distributed, we can write (akin to Equation 7.25):

$$\text{corr}\{Y_l, Y_{l'}\} = \frac{(Y_l - \overline{Y}_l)^t(Y_{l'} - \overline{Y}_{l'})}{\sigma_{Y_l}\sigma_{Y_{l'}}} \approx \frac{\sigma'}{\sigma_{Y_l}\sigma_{Y_{l'}}} F^t_l F_{l'} \tag{7.28}$$

where $\overline{Y}_l$ is the average production of location l, σ_{Y_l} is the standard deviation of Y_l, σ' is the

diagonal of the covariance matrix ($(B-\overline{B})^t(B-\overline{B}) \approx \sigma' I_{N_f}$). We renormalize the correlation to build a similarity metric between 0 and 1 by adding 1 and dividing by 2. Hence, we can estimate a measure of the similarity between the factor endowments of two locations, l and l' as:

$$\phi_{ll'} = (1 + \text{corr}\{Y_l, Y_{l'}\})/2 \tag{7.29}$$

where corr represents the Pearson correlation between vectors, Y_l and $Y_{l'}$. Following Bahar et al. (2014), we refer to this location-location similarity matrix as the country space.

7.2.3 Scaling the matrices

Locations and industries differ greatly in size. It is often useful to normalize each location and each industry using, for example, the revealed comparative advantage (Balassa, 1964) or location quotient or the relative per capita output of each industry in each location. We can show that the correlations calculated over the normalized data have the same information regarding the input similarity of industries or the endowment similarity of locations. To show this, let us assume that we divide each industry by its relative size, s_i, and each location by its corresponding size, s_l. We define the $\widehat{Y}$, $\widehat{A}$ and $\widehat{F}$ matrices such that $\widehat{Y}_{il} = Y_{il}/(s_i s_l)$, $\widehat{A}_{fi} = s_i A_{fi}$ and $\widehat{F}_{fl} = F_{fl}/s_l$ then

$$\widehat{A}\widehat{Y} = \widehat{F} \tag{7.30}$$

All the previous results will follow in this renormalized space.

Unfortunately, for the world as a whole we do not have the production data for each industry in each country. The closest data source that we can readily obtain is data on country exports. Here we will show how by using the normalized version of the export dataset we can obtain a very good approximation to their production correlation counterparts. Production is the sum of locally consumed and exported portions of outputs of industries in that location. Mathematically, we can write this as:

$$Y_{il} = X_{il} + C_{il} \tag{7.31}$$

where X_{il} represents net exports and C_{il} represents local consumption. Subtracting the mean output of the industry i in all locations we obtain:

$$Y_i - \overline{Y}_i = (X_i - \overline{X}_i) + (C_i - \overline{C}_i) \tag{7.32}$$

Assuming homothetic preferences worldwide, and normalizing each industry element by its

size, we can assume that $C_i = \overline{C}_i$. Therefore, correlations of columns of Y can be inferred from correlations of columns of X. Similarly, we can also look at the column vectors of Y and X:

$$Y_l - \overline{Y}_l = (X_l - \overline{X}_l) + (C_l - \overline{C}_l) \tag{7.33}$$

Again, assuming homothetic preferences worldwide, and normalizing each location by its size then each country would consume the same share of products, implying that $C_l = \overline{C}_l$. Consequently, correlations between the columns of Y can be inferred from the correlations between the columns of X.

7.2.4 Simulating the estimators on an HOV toy model

We test the effectiveness of our estimators of Y_{il} by creating a surrogate dataset using a toy model based on our HOV model. First, we verify that our industry similarity index captures the distance between the factor requirements of industries, and that our location similarity index captures the distance between the factor endowments of locations. Next, we estimate how well our density measures predict the output of each industry-location. We will then study the impact of different neighborhood filters at different levels of noise.

To create our surrogate dataset, we set the number of industries N_i and the number of locations N_l both equal to 200. We also set the number of factors N_f equal to 200 to ensure that the A matrix is invertible. We then populate the A and F matrices using a uniform random distribution with values between zero and one. From these factor requirement and endowment matrices, we can produce a 200 by 200 matrix of output values Y_{il} using the equation $Y = A^{-1}F$.

We can now explore whether the correlation between pairs of Y rows is related to the correlation between pairs of A^{-1} rows, meaning that the similarity of production or export intensity of products across all locations carries information about the similarity of their factor requirements, as indicated by Equation 7.25. We randomly select 5,000 A^{-1} and F matrices and test the validity of this equation. We note that the random selection of both matrices simultaneously puts no inherent structure into these matrices and in reality we expect to observe more structures matrices. Even in the random case, the correlation between the actual and estimated numbers exhibit is 0.532 ∓ 0.014. We also test whether the correlation between pairs of columns of Y is related to the correlation between the corresponding columns of factor endowments F as suggested by Equation 7.28 and obtained the same correlation coefficient. These results confirm that the correlations of rows (columns) in the Y matrix are informative about the correlation between rows in the A^{-1} matrix

(columns in the F matrix). When we put more structure into the model by introducing higher order correlations in the A^{-1} matrix or the F matrix, our correlation coefficients increase significantly.

Next, we use our density index to estimate the intensity of output of each industry-location cell. To do this, we estimate the product space density of industry i in location l by calculating the weighted average of the intensities of the k most similar products in location l with the weights being the similarity coefficients of each industry to industry i. We also calculate the country space density of industry i in location l by estimating the weighted average of the intensity of industry i across the k most similar locations. Setting $k = 50$ and iterating the simulation through 5,000 trials, we find that our hybrid density model (i.e., a regression including both industry density and location density) is a powerful predictor of industry-location output (mean $R^2 = 0.784$, with 95% confidence interval of 0.715 0.853 across all simulations). However, we need not fix the neighborhood filter at $k = 50$. In Figure 1, the uppermost line shows the effect of neighborhood size on the R^2. We see that the highest R^2 value is found at $k = 4$.

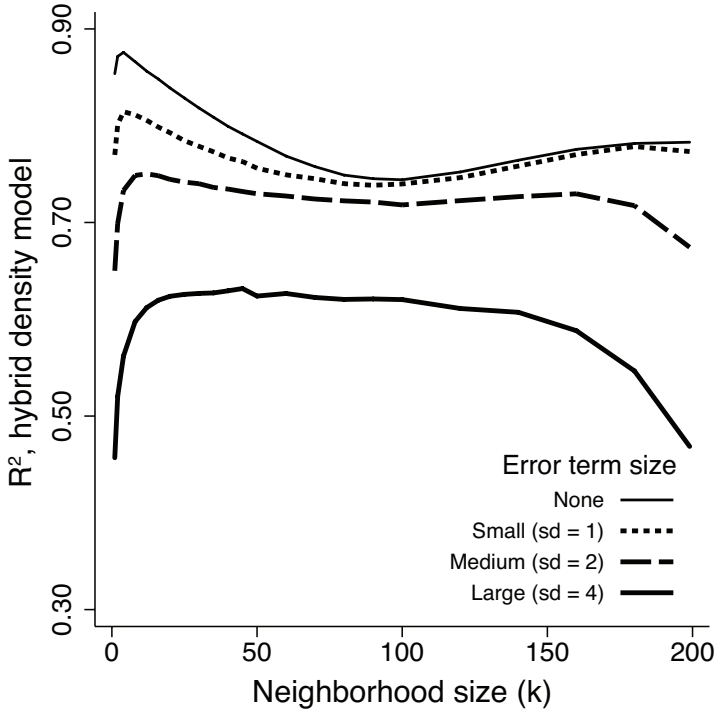

Figure 3: *Simulation of association between underlying output and hybrid density model, by size of neighborhood and noise level.*

Finally, we can extend our simulation to examine the effect of noise in the observed output. Beginning with the $Y = A^{-1}F$ used above, suppose that observed output, $\tilde{Y}_{il}$, is

affected by a random error term, ε_{il}, with a normal distribution around a mean of zero:

$$\tilde{Y}_{il} = Y_{il} + \varepsilon_{il} \tag{7.34}$$

Because the error term is not correlated across location or industry, we can expect that averaging our density index over several neighbors will reduce the effect of noise on our results. That is, we can achieve a better estimate of the noise-free output Y_{il} by averaging the observed, noisy output $\tilde{Y}_{il}$ of the most similar industries and locations, since the error in their output levels might cancel out. Our simulations confirm this hypothesis. We test three levels of noise in the output. Given that the standard deviation of Y_{il} in our surrogate data is 1.994 (median value from 5,000 trials) we use assign the noise term standard deviations equal to 1, 2 and 4, which are approximately half, equal to and double the standard deviation of Y_{il}, respectively.

In Figure 1, we see the effect of increasing the size of the error term on the correlation between the density variables and the actual product intensity. First, we note that, as expected, a larger error term does reduce the R^2 of our estimates, though the decline is relatively small. Second, as noise increases, the R^2 peak tends to move toward mid-range k values, suggesting that the tradeoff between focusing on more related industries and averaging over a broader set of observations moves in favor of the latter. At the same time, the relationship between k and R^2 levels out as noise increases. For example, with a noise level of 2, the R^2 curve is fairly flat with predictive power roughly equal between k values of 4 and 150. This result suggests that finding the optimal neighborhood size may not be a first-order concern for our empirical tests.

References

Antweiler, Werner and Daniel Trefler, "Increasing Returns and All That: A View from Trade," *The American Economic Review*, 2002, *92* (1), 93–119.

Bahar, Dany, Ricardo Hausmann, and César A. Hidalgo, "Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?," *Journal of International Economics*, 2014, *92* (1), 111 – 123.

Balassa, Bela, "The purchasing-power parity doctrine: a reappraisal," *The Journal of Political Economy*, 1964, *72* (6), 584–596.

Bowen, Harry P, Edward E Leamer, and Leo Sveikauskas, "Multicountry, multifactor tests of the factor abundance theory," *The American Economic Review*, 1987, *77* (5), 791–809.

Bustos, Sebastián, Charles Gomez, Ricardo Hausmann, and César A Hidalgo, "The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems," *PloS one*, 2012, *7* (11), e49393.

Conway, Patrick J, "The case of the missing trade and other mysteries: Comment," *The American Economic Review*, 2002, *92* (1), 394–404.

Costinot, Arnaud and Dave Donaldson, "Ricardo's Theory of Comparative Advantage: Old Idea, New Evidence," *The American Economic Review*, 2012, *102* (3), 453–458.

_ , _ , and Ivana Komunjer, "What goods do countries trade? A quantitative exploration of Ricardo's ideas," *The Review of Economic Studies*, 2012, *79* (2), 581–608.

Davis, Donald R and David E Weinstein, "An Account of Global Factor Trade," *The American Economic Review*, 2001, *91* (5), 1423–1453.

_ , _ , Scott C Bradford, and Kazushige Shimpo, "Using International and Japanese Regional Data to Determine When the Factor Abundance Theory of Trade Works," *The American Economic Review*, 1997, *87* (3), 421–46.

Deardorff, Alan V, "The general validity of the Heckscher-Ohlin theorem," *The American Economic Review*, 1982, *72* (4), 683–694.

_ , "Testing trade theories and predicting trade flows," *Handbook of international economics*, 1984, *1*, 467–517.

Debaere, Peter, "Relative factor abundance and trade," *Journal of Political Economy*, 2003, *111* (3), 589–610.

Delgado, Mercedes, Michael E Porter, and Scott Stern, "Clusters and entrepreneurship," *Journal of Economic Geography*, 2010, *10* (4), 495–518.

_ , _ , and _ , "Clusters, convergence, and economic performance," Technical Report, National Bureau of Economic Research 2012.

Dornbusch, Rudiger, Stanley Fischer, and Paul Anthony Samuelson, "Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods," *The American Economic Review*, 1977, *67* (5), 823–839.

Eaton, Jonathan and Samuel Kortum, "Technology, geography, and trade," *Econometrica*, 2002, *70* (5), 1741–1779.

Ellison, Glenn and Edward L Glaeser, "The geographic concentration of industry: does natural advantage explain agglomeration?," *The American Economic Review*, 1999, *89* (2), 311–316.

_ , _ , and William R Kerr, "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *The American Economic Review*, 2010, *100* (3), 1195–1213.

Feenstra, Robert C, *Advanced international trade: theory and evidence*, Princeton University Press, 2003.

Gaulier, Guillaume and Soledad Zignago, "BACI: International Trade Database at the Product-level The 1994-2007 Version," 2010. CEPII Working Paper, No: 2010-23.

Glaeser, Edward L, Hedi D Kallal, José A Scheinkman, and Andrei Shleifer, "Growth in Cities," *Journal of Political Economy*, 1992, pp. 1126–1152.

Hakura, Dalia S, "Why does HOV fail?: The role of technological differences within the EC," *Journal of International Economics*, 2001, *54* (2), 361–382.

Hausmann, Ricardo and Bailey Klinger, "Structural Transformation and Patterns of Comparative Advantage in the Product Space.," 2006. Center for International Development at Harvard University.

_ and _ , "The structure of the product space and the evolution of comparative advantage," 2007. Center for International Development at Harvard University.

_ , César A Hidalgo, Sebastián Bustos, Michele Coscia, Sarah Chung, Juan Jimenez, Alexander Simoes, and Muhammed A Yildirim, *The Atlas of Economic Complexity: Mapping Paths to Prosperity*, Puritan Press, 2011.

Helpman, Elhanan and Paul R Krugman, *Market structure and foreign trade: Increasing returns, imperfect competition and the international economy*, The MIT press, 1985.

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

Leamer, Edward E, "The Leontief Paradox, Reconsidered," *Journal of Political Economy*, 1980, *88* (3), 495–503.

Leontief, Wassily, "Domestic production and foreign trade; the American capital position re-examined," *Proceedings of the American Philosophical Society*, 1953, *97* (4), 332–349.

_ , "Factor proportions and the structure of American trade: further theoretical and empirical analysis," *The Review of Economics and Statistics*, 1956, *38* (4), 386–407.

Linden, Greg, Brent Smith, and Jeremy York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, 2003, *7* (1), 76–80.

Maskus, Keith E and Shuichiro Nishioka, "Development-related biases in factor productivities and the HOV model of trade," *Canadian Journal of Economics/Revue canadienne d'économique*, 2009, *42* (2), 519–553.

Neffke, Frank, Martin Henning, and Ron Boschma, "How do regions diversify over time? Industry relatedness and the development of new growth paths in regions," *Economic Geography*, 2011, *87* (3), 237–265.

Porter, Michael, "The economic performance of regions," *Regional Studies*, 2003, *37* (6-7), 545–546.

Porter, Michael E, *On competition*, Harvard Business Press, 2008.

Reimer, Jeffrey J, "Global production sharing and trade in the services of factors," *Journal of International Economics*, 2006, *68* (2), 384–408.

Resnick, Paul, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in "Proceedings of the 1994 ACM conference on Computer supported cooperative work" ACM 1994, pp. 175–186.

Ricardo, David, *On the Principles of Political Economy and Taxation*, John Murray, London, 1817.

Rodrik, Dani, "Unconditional convergence in manufacturing," *The Quarterly Journal of Economics*, 2013, *128* (1), 165–204.

Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," in "Proceedings of the 10th international conference on World Wide Web" ACM 2001, pp. 285–295.

Tolbert, Charles M and Molly Sizer, "US commuting zones and labor market areas: A 1990 update," 1996. Economic Research Service Staff Paper 9614.

Trefler, Daniel, "International Factor Price Differences: Leontief Was Right!," *Journal of Political Economy*, 1993, *101* (6), 961–87.

_ , "The Case of the Missing Trade and Other Mysteries," *The American Economic Review*, 1995, *85* (5), 1029–1046.

_ **and Susan Chun Zhu**, "Beyond the algebra of explanation: HOV for the technology age," *The American Economic Review*, 2000, *90* (2), 145–149.

_ **and** _ , "The structure of factor content predictions," *Journal of International Economics*, 2010, *82* (2), 195–207.

Vanek, Jaroslav, "The Factor Proportions Theory: The N-Factor Case," *Kyklos*, 1968, *21* (4), 749–756.

Zhu, Susan Chun and Daniel Trefler, "Trade and inequality in developing countries: a general equilibrium analysis," *Journal of International Economics*, 2005, *65* (1), 21–48.