

# Implied Comparative Advantage

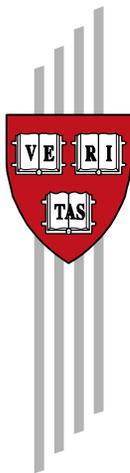
Ricardo Hausmann, César A. Hidalgo, Daniel P. Stock, and  
Muhammed A. Yildirim

CID Faculty Working Paper No. 276

January 2014

*Revised September 2019*

© Copyright 2019 Hausmann, Ricardo; Hidalgo, César A.; Stock,  
Daniel P.; Yildirim, Muhammed A; and the President and Fellows of  
Harvard College



## Working Papers

Center for International Development  
at Harvard University

# Implied Comparative Advantage <sup>\*</sup>

Ricardo Hausmann      César A. Hidalgo      Daniel P. Stock  
Muhammed A. Yıldırım<sup>†</sup>

September 2019

## Abstract

The comparative advantage of a location shapes its industrial structure. Current theoretical models based on this principle do not take a stance on how comparative advantages in different industries or locations are related with each other, or what such patterns of relatedness might imply about the underlying process that governs the evolution of comparative advantage. We build a simple Ricardian-inspired model and show this hidden information on inter-industry and inter-location relatedness can be captured by simple correlations between the observed patterns of industries across locations or locations across industries. Using the information from related industries or related locations, we calculate the *implied comparative advantage* and show that this measure explains much of the location's current industrial structure. We give evidence that these patterns are present in a wide variety of contexts, namely the export of goods (internationally) and the employment, payroll and number of establishments across the industries of subnational regions (in the US, Chile and India). The deviations between the *observed* and *implied* comparative advantage measures tend to be highly predictive of future industry growth, especially at horizons of a decade or more; this explanatory power holds at both the intensive as well as the extensive margin. These results suggest that a component of the long-term evolution of comparative advantage is already implied in today's patterns of production.

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

---

<sup>\*</sup>We thank Philippe Aghion, Pol Antràs, Sam Asher, Jesus Felipe, Elhanan Helpman, Asim Khwaja, Paul Novosad, Andrés Rodríguez-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.

<sup>†</sup>Hausmann: Center for International Development at Harvard University (CID), Harvard Kennedy School & Santa Fe Institute. Hidalgo: ANITI Chair at the University of Toulouse, Stock: Carnegie Mellon University & CID, Yıldırım: Koç University & CID. Emails: ricardo\_hausmann@harvard.edu (Hausmann), hidalgo@mit.edu (Hidalgo), dstock@andrew.cmu.edu (Stock), mayildirim@ku.edu.tr (Yıldırım).

David Ricardo (1817)'s seminal theory predicts that locations specialize in the goods in which they have a comparative advantage, meaning that they enjoy a higher relative productivity. Yet these comparative advantage patterns are not random, nor are they set in stone; theories detailing the evolution of a locations' productivities date back to the work of Marshall (1890) more than a century ago. Since then, there have been many studies highlighting the importance of sectoral or regional relatedness in the evolution of comparative advantage or the production structures of regions. Here, we take a complementary stance, giving evidence that the relatedness of industries and locations could be used to develop a measure of counterfactual or *implied comparative advantage*, and that such a measure helps explain how locations' comparative advantage change over time.

Imagine the following thought experiment. You have conducted the first-ever economic census of your country, containing the output matrix for all cities across all industries. However, due to some accident, your computer randomly erases a few entries in this matrix. How would you guess what those entries were if you had no information other than the surviving part of the matrix? The current theoretical models would not help you use the surviving data to predict the inter-industry variation in output at the city level for the missing data.

In this paper we extend the neo-Ricardian models by assuming that the industrial relatedness causes comparative advantage levels to be differentially correlated across industries in a manner that can be empirically estimated. This structure implies that the comparative advantage of a location in an industry can be estimated from its comparative advantage in related industries, even for the industries that are currently absent from the location. Furthermore, information about the comparative advantage levels of locations with similar attributes should be informative of the comparative advantage levels of industries in a given location. Hence, we can infer the similarity in the productive orientation of locations from the similarity in their output structure. Symmetrically, the intensity with which a location engages in an industry should be related to the intensity with which it engages in similar industries, where industry similarity is calculated from the pattern of revealed comparative advantage across all locations.

We use these similarity measures to generate proxies of the *implied comparative advantage* of a location in an industry, and show that it is strongly correlated with the *observed* or *revealed comparative advantage* of that country in that industry. In addition, these estimates are helpful in explaining 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 or residual terms of our prediction are not merely noise, but

are actually predictive of future changes. Put it simply, we show that (1) information on hidden endowments and requirements can be recovered from an analysis of the realized economic structures of the locations (*i.e.*, *observed* comparative advantage) and (2) long-term changes in *observed* comparative advantage are associated with the gap between current comparative advantage and the comparative advantage implied by the recovered information.

Given both national and subnational applicability, our approach is inspired and informed by techniques or models used multiple fields, especially trade theory, urban economics, and regional economics. We base our theoretical models on the international trade literature as many recent regional development studies do (*e.g.*, Davis and Dingel (2014); Costinot et al. (2016); Caliendo et al. (2017)). Our empirical work is more intimately related to related diversification literature (Hidalgo et al., 2018; Boschma, 2017). Our results are important in shaping policy discussions: given the special importance of the emergence of new industries, we believe that our results could be used in better informing policies related to this process.

According to the Ricardian theory of trade, the pattern of which location specializes in which goods is determined not by its absolute productivity in that industry but instead by its productivity relative to that of other industries in the same location. Although Ricardo introduced this idea using two countries (England and Portugal) and two products (cloth and wine) almost two centuries ago (Ricardo, 1817), the multi-location 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 approaches infer a location's productivity in a certain industry from its observed pattern of trade and have been successful in explaining a significant portion of bilateral and subnational trade and economic activity. 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.<sup>1</sup> This is an important shortcoming as there are many instances in which it would be useful to infer the productivity level that a location 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. 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. We provide evidence that appears to contradict this.

---

<sup>1</sup>Deardorff (1984), as quoted by Costinot et al. (2012), says that "*The . . . problem is implicit in the Ricardian model itself . . . [because] the model implies complete specialization in equilibrium . . . 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.*"

Essentially, our Ricardian-inspired 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 factor-based explanation given by Heckscher-Ohlin and later extended by Vanek (1968) 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 as well. With a factor-based interpretation, the revealed comparative advantage of a location in an industry can be inferred from its revealed comparative advantage in industries that have similar production functions or locations that have similar factor endowments. Interestingly, our results can be derived without information regarding production functions or factor endowments.

Our findings do not only pertain to international trade: we obtain similar results when we construct measures of industry-location structure using 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 growth: they correlate with future growth rates of industry-locations, as well as with the appearance and disappearance of new industries in each location.

### *Related literature*

This paper is related to several strands of literature given its scope covers both national and subnational data. Specifically, it builds on Hausmann and Klinger (2006), Hidalgo et al. (2007), Hausmann and Klinger (2007) and Bahar et al. (2014) but develops a theoretical framework and explores both the extensive and the intensive margins of industrial evolution of regions. Building on these studies, Boschma and Capone (2015) analyze the interaction between relatedness and institutions and finds that different varieties of capitalism result in different diversification patterns. Petralia et al. (2017) find that the related diversification is also important at the technological development of countries especially at initial stages of development. Boschma et al. (2012, 2013) apply a similar approach to understand the regional diversification in Spain. Neffke et al. (2011) show that regions diversify into related industries, using an industry relatedness measure based on the co-production of products within plants.

These studies could be thought as a part of larger relatedness literature (Hidalgo et al.,

2018; Boschma, 2017). Relatedness measures have been used to understand the relationship between technology intensity of an industry and agglomeration (Liang and Goetz, 2018) and to understand how scientific knowledge diffuses between cities (Boschma et al., 2014) as well.

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. Hanlon and Miscio (2017) further show that the historical pattern of location distribution of industries in Britain are shaped by the agglomerative forces as well. Delgado et al. (2010, 2015) and Porter (2003) use US sub-national data to explain employment growth at the city-industry level, using the presence of related industry clusters. Lu et al. (2016) explores the effect of co-located clusters in the emergence of new clusters and find differential interactions depending on the maturity of the cluster. Implicitly, the observed formation of clusters in a location and the location's comparative advantage are linked with each other. Beaudry and Schiffauerova (2009) surveys the literature to determine whether Marshallian forces or diversity of a region is more effective on the economic progress of regions. Our work does not take a stance in that regard but the relatedness measures that we use capture more than the Marshallian forces.

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 user's 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.

Our paper 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 Fréchet 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. Previous Ricardian literature, however, cannot infer relative productivities of industries that do not exist. An exception is Costinot et al. (2016) where they estimate implied or counter-factual productivity parameters for agricultural industries using agroeconomic 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.

Here, we take the two dimensional industry-location matrices and predict the evolution of the productivity based on our *implied* comparative advantage metric. The economic complexity literature based on Hidalgo and Hausmann (2009) creates one-dimensional projections from the same matrix and develops metrics to quantify country complexity and product sophistication. This work inspired different metrics such as the country fitness and product quality metrics developed in Tacchella et al. (2012, 2013), Caldarelli et al. (2012), Cristelli et al. (2013) and Bustos and Yildirim (2019a). These measures can also be used to model new product appearances in the context of evolution of the complexity. Nevertheless, they do not aim to predict industry-location level specification.

This paper is structured as follows. Section 1 derives provides the theoretical motivation of our results. Section 2 discusses the data and the methodology to build our variables of interests. Section 3 presents our main empirical results for the intensive margin, the growth of industries in location and the extensive margin. In Section 4, we test some direct implications of our model, and Section 5 evaluates the alternative explanations and robustness of our results. In Section 6, we conclude with a discussion of the implications of our findings.

## 1 Theoretical motivation

In this section, we use a modified Ricardian framework to show how the similarities between the output of industries across different locations (and similarities between the output of locations across different industries) can contain information on the *true* comparative advantage of a location, *i.e.*, the hidden match between the requirements of industries

and the ability of locations to meet those requirements. As we argue in the Introduction, a 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 the co-location patterns of the same industry across countries. However, it is possible to make a Ricardian model compatible with these observations by incorporating the assumptions that products differ in their technological relatedness, and that 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. By the same token, it would also export that product with an intensity that is similar to that with which it exports technologically related products. Below, we will only focus on the determinants of the comparative advantage. Alternatively a general equilibrium model can be written and the conclusions are consistent with our predictions. In the Appendix, we also derive measures that capture the similarity between industries and between locations using a factor-based model like Heckscher-Ohlin-Vanek (Vanek, 1968) model.

Here, we will construct a particular relation between the requirements of an industry and endowments of a location, showing how their interaction might effect the observed global patterns of output, and what can be inferred from these patterns. Let's denote the size (*e.g.*, output, payroll, employment, or number of establishments) of an industry  $i$  in a location  $l$  with  $y_{il}$ . Suppose the total industry size in the sample universe (which is the world if we are dealing with a national data or the country if we are dealing with the subnational data) is  $Y_i \equiv \sum_l y_{il}$ . The location's expected size will depend on the size of the location relative to size of the sample universe. Let's denote size of the location with  $W_l$  and size of the sample universe with  $W \equiv \sum_l W_l$ . If the industries were to be distributed relative to the size of a location, we would expect size of industry  $i$  in location  $l$  to be:

$$\hat{y}_{il} = Y_i \frac{W_l}{W}.$$

However, in real life, this distribution is not based on the size. Instead, it follows the pattern of the comparative advantage. We can develop a measure of comparative advantage by dividing the observed size of the industry to its expected size to capture how more likely it is for a location to host the industry:

$$r_{il} = \frac{y_{il}}{\hat{y}_{il}}.$$

By taking logs, using the definition of the expected size and re-arranging the terms we arrive at:

$$\log(y_{il}) = \log(r_{il}) + \log(Y_i) + \log(W_l) - \log(W) \quad (1.1)$$

In particular, if  $y_{il}$  is the exports of country  $l$  in industry  $i$ , and  $W_l$  is the total exports of the country, then  $r_{il}$  becomes Balassa (1964)'s Revealed Comparative Advantage (RCA) measure. Instead, if we use  $W_l$  to be the population of the location, we arrive at the Revealed per-Capita Advantage (RpCA), which will be more explicitly defined in Section 2.2. Alternatively, if we use  $y_{il}$  to be the number of employees in industry  $i$  in location  $l$  and  $W_l$  as the total employment in location  $l$  we arrive at the Location Quotient (LQ) measure that has been widely used. Similarly, we can also use  $y_{il}$  to be the total payroll in industry  $i$  in location  $l$ , and  $W_l$  as the total payroll in location  $l$  and come up with a new measure of comparative advantage.

In a sense, Equation 1.1 is a decomposition of the size of an industry in a location. It has a component that captures the dynamics in the total size of the industry ( $Y_i$ ), another component that captures the location specific dynamics ( $W_l$ ), and a portion that is location-industry specific ( $r_{il}$ ). In our empirical analysis, we will be focusing in this interaction term. Normalizing output values in this way is attractive: it lets us strip out the scaling effects that exist purely at the location level (*e.g.*, the population size of a country or total exports of a country) and the industry level (*e.g.*, the global demand for a commodity), and instead focus on explaining the interplay between industries and locations. That is, instead of asking questions like “Why is employment growth higher in Boston than in Kansas City?” or “Why is employment in retail services growing faster than electronics manufacturing?” we ask questions in the class of “Why is electronics manufacturing growing relatively faster in Boston than in Kansas City?”

## 1.1 Modeling comparative advantage

Having defined our measure of industry-location intensity, we can now build a framework to model how these intensities are generated. We will assume that the efficiency with which industry  $i$  functions in location  $l$  depends on the distance between the requirements of industry  $i$  and endowments of location  $l$ ; specifically, we measure this distance in a compact and convex metric subspace in  $\mathbb{R}^n$  denoted by  $\mathcal{S}$ . Suppose the requirements of the industry  $i$  are characterized by a parameter  $\psi_i \in \mathcal{S}$  and the endowments of location  $l$  is characterized by a parameter  $\lambda_l \in \mathcal{S}$ . The output intensity of industry  $i$  in location  $l$

$(r_{il})$  will depend on some function of the distance between  $\psi_i$  and  $\lambda_l$ . More concretely,

$$r_{il} = f(d(\psi_i, \lambda_l)) \quad (1.2)$$

where  $d$  is the distance metric on our compact metric space  $\mathbb{S}$ , and  $f$  is a strictly decreasing function of that distance, such that  $f(0) = 1$  and  $f(d_{max}) = 0$ , where  $d_{max}$  is the maximum distance in  $\mathbb{S}$  – *i.e.*, output intensity is increasingly high as the industry requirement and location endowment are closer together.

In reality, we are not able to observe  $\psi_i$  and  $\lambda_l$  directly – they are hidden characteristics. However, we do observe the  $r_{il}$ , and can in fact use them to glean information about the underlying  $\psi_i$  and  $\lambda_l$  values. For example, it is clear that the difference between a location's comparative advantage in two industries,  $r_{il}$  and  $r_{i'l}$ , is an increasing function of the distance between the  $\psi_i$  and  $\psi_{i'}$ . By the same token, the difference in the intensity of output of the same industry across two locations,  $r_{il}$  and  $r'_{il}$ , would be an increasing function of the difference in the  $\lambda_l$  and  $\lambda_{l'}$ .

We can generalize this intuition to incorporate information from all of the industry-location pairs. To do so, we must first construct a metric that lets us formally relate one industry to another (by comparing their intensities in the same locations) or one location to another (by comparing their intensities in the same industries). Suppose we start with the normalized output intensity  $r_{il}$  for each industry in each location. We can calculate a 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<sup>2</sup>  $r_i$  and  $r_{i'}$  across all locations:

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 \quad (1.3)$$

Symmetrically, we define the country space similarity matrix  $\phi_{ll'}$  between two locations  $l$  and  $l'$  as the scaled Pearson correlation between  $r_l$  and  $r_{l'}$  across all industries:

$$\phi_{ll'} = (1 + \text{corr}\{r_l, r_{l'}\})/2 \quad (1.4)$$

We want to show that  $\phi$  captures the distance between industries or the distance between locations in our compact space  $\mathbb{S}$ . In particular, without having the knowledge of what  $\psi_i$  and  $\psi_{i'}$  are, we would like to infer the distance between them using  $\phi_{ii'}$ . In fact, we can prove (in Section A.2.3 of Appendix) that if we model our space  $\mathbb{S}$  as a unit-sphere

---

<sup>2</sup>As a notational convention, when we skip an index, it means that it becomes a vector over the skipped variable. In this case,  $r_i$  is the vector whose elements are  $r_{il}$ .

in an arbitrary number of dimensions ( $\mathbb{R}^n$ ), and use any arbitrary function  $f$  that satisfies the rules described above, then it follows that  $\phi_{ii'}$  strictly decreases as the distance between  $\psi_i$  and  $\psi_{i'}$  increases:

$$\frac{d\phi_{ii'}}{dd(\psi_{i'}, \psi_i)} < 0$$

and  $\phi_{ll'}$  strictly decreases as the distance between  $\lambda_l$  and  $\lambda_{l'}$  increases:

$$\frac{d\phi_{ll'}}{dd(\lambda_{l'}, \lambda_l)} < 0.$$

In other words, for any two industries (or locations), the closer their hidden requirements (or endowments) are, the more highly correlated their observable comparative advantages will be across locations (or across industries).

Going forward, we will focus on the simplest case, where  $\psi$  and  $\lambda$  are points on a unit-sphere of one dimension, *i.e.*, the unit circle, which we denote as  $\mathbb{U}$ .<sup>3</sup> On  $\mathbb{U}$ , output intensity will be maximized when  $\psi_i = \lambda_l$ ; in the opposite case, where  $\psi_i$  and  $\lambda_l$  are on antipodal sides of the circle (and distance is 0.5), output would be zero. While our expected outcomes hold for any function  $f(d(\psi_i, \lambda_l))$  (following the form described above), we will now assign a specific functional form for  $f$ , in order to calculate actual  $r_{il}$  values:

$$f(d(\psi_i, \lambda_l)) = 1 - 4d^2(\psi_i, \lambda_l). \quad (1.5)$$

If we assume that  $\psi_i$  and  $\lambda_l$  are uniformly distributed on the unit circle,  $\mathbb{U}$ , 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  $\psi_i$ s (see Appendix for the details of the calculation):

$$\phi_{ii'} = 1 - 15 \left( d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'}) \right)^2 \quad (1.6)$$

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

$$\phi_{ll'} = 1 - 15 \left( d(\lambda_l, \lambda_{l'}) - d^2(\lambda_l, \lambda_{l'}) \right)^2 \quad (1.7)$$

Note that for distance  $d = 0$ , the expected proximity would be 1. If distance is equal to its maximum value ( $d = 1/2$ ) then the expected proximity would be the minimum.<sup>4</sup>

<sup>3</sup>We chose the unit circle rather than a line to avoid boundary effects of the space. For instance, for an interval like  $[0, 1]$ , the boundaries, 0 and 1, will introduce break points.

<sup>4</sup>The minimum value of the similarity turns out to be 1/16. This corresponds to a correlation of -7/8. The reason that we do not observe a correlation of -1 is that even at the antipodal locations, it is impossible

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.

## 1.2 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. We can imagine an industry  $i$  in location  $l$ , where we do not know the comparative advantage  $r_{il}$  and thus wish to estimate it. What we do know is that whenever we observe industry  $i$  in other locations, it is produced in virtually identical intensities to a second industry,  $i'$ ; it follows that  $\phi_{ii'} \approx 1$ . Likewise, when we look at the other industries in location  $l'$ , we note that their intensities are virtually identical to those of the industries in a second location,  $l''$ ; this means that  $\phi_{ll''} \approx 1$ . Based on equations 1.6 and 1.5, we know that  $\phi_{ii'} \approx 1 \Rightarrow \psi_i \approx \psi_{i'}$  and  $\phi_{ll''} \approx 1 \Rightarrow \lambda_l \approx \lambda_{l''}$ . Plugging these into our formula for  $r_{il}$  in Equation 1.6 would imply that  $\phi_{ii'} \approx 1 \Rightarrow r_{i'l} \approx r_{il}$  and  $\phi_{ll''} \approx 1 \Rightarrow r_{il''} \approx r_{il}$ . That is, if we can find a nearly identical comparator industry (based on its intensity in other locations), then our measure of *implied* comparative advantage is the intensity of that industry in the same location. Likewise, if we find a nearly identical comparator location (based on its intensity across other industries), then we can use the intensity of the same industry in that location as our *implied* comparative advantage proxy.

In practice, however, we argue that the best estimate of *implied* comparative advantage comes not from a single comparator; instead, a weighted mean of the top  $k$  most similar comparators is often more accurate (Sarwar et al., 2001). We can use our model to illustrate two ways in which such an approach would yield better results.

First, we find that there are no perfect comparators in the real world – it is exceedingly rare to find nearly identical pairs of industries or locations in our datasets.<sup>5</sup> Thus, if we base our proxy on the single most related industry or location, we may be introducing a large error whenever  $\phi_{ii'} < 1$  or  $\phi_{ll''} < 1$ . In fact, our model implies that the most similar industry or location is not necessarily the most accurate proxy. Take, for example, a location  $l$  with two highly similar comparators,  $l'$  and  $l''$ . In cases where  $\lambda_{l'} > \lambda_l > \lambda_{l''}$

---

to get perfect anti-correlation on the unit circle. Even in that case, the points at  $90^\circ$  and  $270^\circ$  will have similar values of comparative advantage.

<sup>5</sup>Section 2.2 below analyzes the distribution of similarity values in real-world data. For most locations, the best comparator location is far from identical (median similarity = 0.69). Industries, however, are likelier to have a highly similar best comparator (median = 0.92).

or  $\lambda_{i''} > \lambda_l > \lambda_{i'}$ , the average of  $r_{i'l'}$  and  $r_{i'l''}$  will be closer to  $r_{il}$  than either  $r_{i'l'}$  or  $r_{i'l''}$  individually.<sup>6</sup> Of course, since we do not observe the  $\lambda$  values in the real world, it may make sense to consider a larger number of comparators; this may improve the odds that we are including locations whose endowments deviate from  $\lambda_l$  in opposite directions, *i.e.*, places whose differences from  $l$  will *cancel out* each other.

Second, it is unlikely that can ever witness the *true* comparative advantage of an industry-location. It is more plausible to imagine a gap between an industry-location's *true* comparative advantage and its *observed* comparative advantage. We can express this in our model by incorporating an additive disturbance term:

$$\tilde{r}_{il} = r_{il} + \varepsilon_{il}, \tag{1.8}$$

Here, the underlying or *true* comparative advantage of an industry,  $r_{il}$ , is determined as before, solely by the distance between the technological requirement of the industry  $\psi_i$  and the technological ability of the location  $\lambda_l$ . Let us assume that the comparative advantage of each industry-location has deviated from this underlying value because of a disturbance term  $\varepsilon_{il,t_0}$ . What could cause this disturbance term? The location might have the potential to be productive in this industry, but has not yet allocated necessary resources to this industry; in this case the disturbance term would be negative. Conversely, a location might have gambled on an industry with a low underlying comparative advantage; it could allocate resources and achieve some level of production in the short term, but would struggle to sustain it. The disturbance term in this case would be positive. As a result of the disturbance term, we no longer see the *true* comparative advantage  $r_{il}$  but instead see only the current or *observed* comparative advantage,  $\tilde{r}_{il,t_0}$ . Again, our comparators will come to rescue: by integrating information from several meaningful comparators, we might be able to predict the *true* comparative advantage. Since we do not observe the sign of the disturbance term, we may want to include a larger number of comparators, to ensure that there are enough positively biased values to offset the negatively biased ones (or vice versa).

We have thus far presented two arguments for incorporating multiple comparators into our measure of comparative advantage. However, we should also note that we expect a tradeoff to arise. As we cast a wider net – *e.g.*, moving from a location's 5 most similar locations to its 25 most similar locations – the additional comparators will become increasingly dissimilar from the original location, diminishing the accuracy of the

---

<sup>6</sup>For example, if we set  $\{\lambda_l, \lambda_{i'}, \lambda_{i''}\}$  to  $\{0.75, 0.77, 0.72\}$  and  $\psi_i$  equal to 0.44, then we get  $\{r_{il}, r_{i'l'}, r_{i'l''}\} = \{0.62, 0.56, 0.69\}$ ; thus, the mean of  $r_{i'l'}$  and  $r_{i'l''} = 0.63$  – is much closer to  $r_{il}$  than either  $r_{i'l'}$  or  $r_{i'l''}$  are.

*implied* comparative advantage measure. Thus, we expect to see an inverted U-shaped relationship between the number of comparators used ( $k$ ) and the accuracy of our weighted average.

Going forward, we can formally define our expected value of the true  $r_{il_0}$  as the weighted mean of the *observed* comparative advantage of the  $k$  nearest neighbors  $\tilde{r}_{il_1}, \tilde{r}_{il_2}, \tilde{r}_{il_3}, \dots, \tilde{r}_{il_k}$ , where both the *nearness* and the weights are given by the similarity parameters  $\phi_{ll'}$ .<sup>7</sup> We refer to this proxy for the *implied* comparative advantage as the country space density:

$$\hat{r}_{il}^{[L]} = \sum_{l' \in L_{lk}} \frac{\phi_{ll'}}{\sum_{l'' \in L_{lk}} \phi_{ll''}} \tilde{r}_{il'} \quad (1.9)$$

with the set  $L_{lk}$  defined as:

$$L_{lk} = \{l' | \text{Rank}(\phi_{ll'}) \leq k\} \quad (1.10)$$

We can also build an analogous metric using the industry similarity measure. In this case, the *implied* comparative advantage of an industry in a location would be the weighted mean of the *observed* comparative advantage of the  $k$  most related industries in the same location:

$$\hat{r}_{il}^{[I]} = \sum_{i' \in I_{ik}} \frac{\phi_{ii'}}{\sum_{i'' \in I_{ik}} \phi_{ii''}} \tilde{r}_{i'l} \quad (1.11)$$

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

$$I_{ik} = \{i' | \text{Rank}(\phi_{ii'}) \leq k\} \quad (1.12)$$

We refer to this variable as the product space density.<sup>8</sup> As the two density measures are symmetric, we can expect that they capture some non-overlapping information; that is, the best proxy may be both of them combined.

Finally, following the rule of thumb established by Duda et al. (2012), we set the  $k$  parameter equal to  $\sqrt{N}$ , *i.e.*, the geometric midpoint between including only the most similar comparator ( $k = 1$ ) and including all comparators ( $k = N$ ). However, we later relax this constraint, testing the expected relationship between  $k$  and accuracy directly.

Note that neither  $\hat{r}_{il}^{[I]}$  nor  $\hat{r}_{il}^{[L]}$  contain direct information from  $\tilde{r}_{il}$ , meaning that we are not using an observation to proxy itself.<sup>9</sup> This also means that we can find non-zero

<sup>7</sup>Note that we would also need to estimate the similarity parameters using *observed* comparative advantage values.

<sup>8</sup>Our density variable is in fact quite similar to the product space density first proposed by Hausmann and Klinger (2006), adapted to describe continuous inputs. Our goal here is not to build the best possible density metric, but to show how such a variable can be motivated by an underlying theoretical model.

<sup>9</sup>Strictly speaking,  $\tilde{r}_{il}$  is still included in the calculation of the similarity measures; we later use a cross-

*implied* comparative advantage values for industry-locations that have current observed values of zero ( $\tilde{r}_{il} = 0$ ); such implied values could be interpreted as the likelihood of a new industry-location being born.

### 1.3 Predicting trends towards true comparative advantage

At this point, we have shown that one can construct a proxy for a location's *implied* comparative advantage in an industry. But what can such a measure tell us about industry-location growth?

Above in Equation 1.8, we introduced a disturbance term,  $\varepsilon_{il}$ , that captures locations' tendency to deviate from their *true* comparative advantage, allocating resources sub-optimally. As such, let us assume that these deviations are unsustainable, and will diminish over long periods of time, *i.e.*,  $\lim_{t \rightarrow \infty} \varepsilon_{ilt} = 0$ .<sup>10</sup> This lets us posit that we will be able to observe a trend towards the *true* Ricardian structure – *i.e.*, that as the temporary disturbance shrinks over time, the observed  $\tilde{r}_{il,t}$  values will grow closer to the true  $r_{il}$ . This implies that there should be a positive relationship between changes in  $\tilde{r}_{il,t}$  and the original gap between *true* and *observed* comparative advantage:

$$\tilde{r}_{il,t_1} - \tilde{r}_{il,t_0} = \beta(r_{il} - \tilde{r}_{il,t_0}), \quad \beta > 0$$

We still cannot test this relationship directly, since we do not witness the *true* comparative advantage values. But since we believe that our *implied* comparative advantage measure is a proxy for *true* comparative advantage (or at least captures novel information on *true* comparative advantage), then we can substitute it into the above equation. This would then imply that we can expect to find a (positive) empirical relationship between the present-day gap between *implied* and *observed* comparative advantage ( $\hat{r}_{il,t_0} - \tilde{r}_{il,t_0}$ ) and future changes in *observed* comparative advantage ( $\tilde{r}_{il,t_1} - \tilde{r}_{il,t_0}$ ):

$$\tilde{r}_{il,t_1} - \tilde{r}_{il,t_0} = \beta(\hat{r}_{il} - \tilde{r}_{il,t_0}), \quad \beta > 0 \tag{1.13}$$

We can now highlight the fundamental questions we wish to test: Is our measure of *implied* comparative advantage a better predictor of the underlying *true* comparative advantage than the current, or *observed* comparative advantage? And if so, can the difference between *observed* and *implied* comparative advantage help explain future industry-

---

validation approach in Section 5.4 to verify that excluding  $\tilde{r}_{il}$  from similarity does not impact our results.

<sup>10</sup>To decrease the clutter in our notations, we only use time indices, denoted by  $t$ , when we explicitly look at the time dimension.

location growth, or even the appearance of new (non-zero) industry-locations? To answer these questions, we will begin by building a simulation of our theoretical model, verifying that the model performs as expected when we supply the underlying parameters. Subsequently, we will test our expectations empirically, using a variety of real-world datasets.

## 1.4 Simulating the model

We can now build our simulation to test our expectations based on our theoretical motivation. We set the dimensions of the simulation to  $N_i = 100$  industries by  $N_l = 100$  locations, and assume a uniform distribution of the  $\psi_i$  and the  $\lambda_l$  along the unit circle  $\mathbb{U}$ . We then use these parameters to calculate the  $r_{il}$  values, fixing  $k = \sqrt{N_l} = \sqrt{N_i} = 10$ , as described above. We model  $\varepsilon_{il}$  in Equation 1.8 as normally distributed with zero mean and variance equal to the variance of  $r_{il}$  (denoted by  $\sigma^2$ ) times a parameter  $s$ , the *noise-to-signal ratio* ( $\varepsilon_{il} \sim N(0, s\sigma^2)$ ). In our simulations we vary this noise-to-signal ratio parameter,  $s$ , and set it to a range of six different values, going from 0 (no error) to 1 (equal parts signal and noise) to 4 (four times more noise than signal). We find that the average standard deviation of  $r_{il}$  converges to 0.298, so we use that value for  $\sigma^2$ . From the noisy output levels ( $\tilde{r}_{il}$ ), we build the product space and country space densities. We then measure the predictive power of the PS and CS densities (and the mean of the two) by running linear regressions with the (noiseless) *true* comparative advantage values. This allows us to benchmark the strength of these correlations against similar correlations between the *observed* (noisy) comparative advantage and *true* comparative advantage values.

**Table 1: Mean  $R^2$  of correlations with equilibrium output, for 5,000 simulations**

Noise-to-signal ratio	Correlation between <i>true</i> comparative advantage and			
	<i>observed</i> comparative advantage	<i>implied</i> comparative advantage		
		PS Density	CS Density	Hybrid
0%	1.000	0.991	0.991	0.995
25%	0.941	0.984	0.984	0.990
50%	0.800	0.963	0.963	0.977
100%	0.500	0.880	0.880	0.923
200%	0.200	0.608	0.608	0.700
400%	0.059	0.124	0.124	0.166

We carry out this exercise multiple times (5,000 simulations) and report on the cor-

relation results. First, we can verify the validity of Equations 1.6 and 1.7: on average, the absolute difference between the observed similarity value and the similarity value implied by the equations is less than 5%.<sup>11</sup>

Next, we can simulate the distribution of the similarity index values that arise from our model (See Figure A.1 in Appendix). When the similarities are calculated using low levels of noise ( $s \leq 100\%$ ), their values are roughly symmetric about the midpoint (0.5, *i.e.*, random), with two peaks (*i.e.*, values clustered around low or high similarity). As noise increases, the peaks move closer to the midpoint; at higher levels of noise ( $s \geq 200\%$ ), there is a single peak at the midpoint. This makes sense, since the coefficients of correlating pure uniform noise would take on a normal distribution (centered on the midpoint). Thus, when we start using real-world data, we should expect to find similarity values symmetrically distributed about the midpoint; depending on how *noisy* the real world is, these values will have a single peak (at the midpoint) or twin peaks (roughly equidistant from the midpoint).

At this point, we can use the simulation to test the relation between our measures of *implied*, *observed* and *true* comparative advantage. To begin, we note that the  $R^2$  values are highly consistent across the simulations, with standard errors all less than a tenth of a percentage point. Table 1 gives the mean  $R^2$  values. We can see that the PS and CS densities perform well at nearly all error levels: other than the *noisiest* simulations,  $R^2$  values are quite high (0.61 to 0.99). Most importantly, the density indices also perform well relative to the *observed* comparative advantage term (the  $\tilde{r}_{il}$  values). As expected based on their construction,  $\tilde{r}_{il}$  values tend to correlate well with *true* comparative advantage (the  $r_{il}$  values) when noise is low; in the extreme case when the error term is nonexistent, *observed* comparative advantage is identical to *true* comparative advantage. However, as we increase the noise-to-signal ratio, *observed* comparative advantage becomes an increasingly weak correlate of *true* comparative advantage. The explanatory power of the PS and CS densities also decreases with increasing noise, but at a much slower rate; at  $s = 100\%$ , the densities are still strongly associated with *true* comparative advantage ( $R^2 = 0.88$ ). This confirms our prediction: in a *noisy* world, where industry-locations are far from their *true* comparative advantages, the *implied* comparative advantage measures may be a better predictor of the underlying values than the observed values. For an empirical setting, the implication is that even if the underlying Ricardian dynamics have a relatively low explanatory power today, our density measures are highly effective at separating their

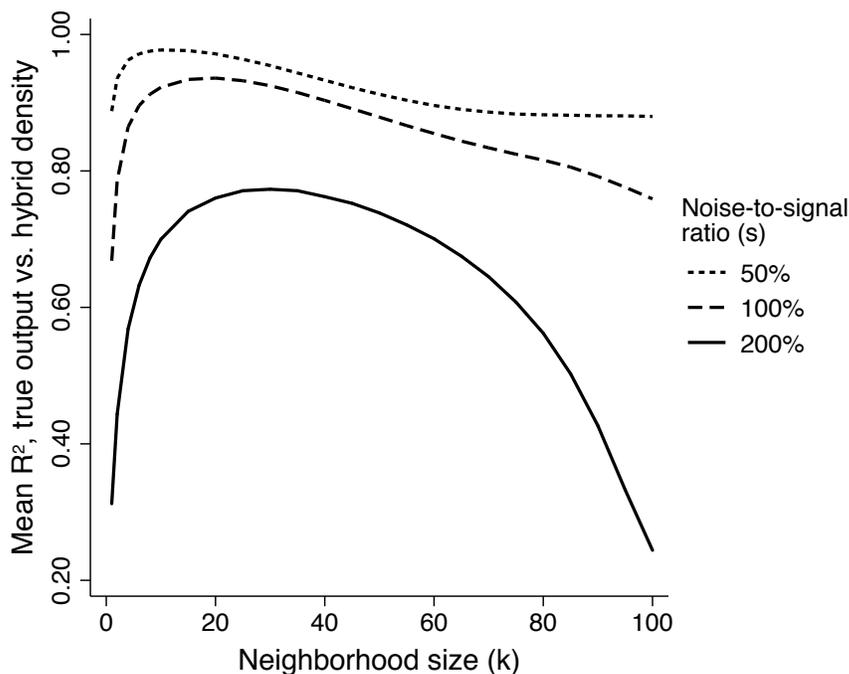
---

<sup>11</sup>The estimated values will never be identical to the values implied by the proof, since the correlations are taken on a finite number of random variables. In fact, the error decreases (to 1.4%) when we increase  $N$  from 100 to 1000.

message from the noise.<sup>12</sup>

Finally, note that the explanatory power of the PS and CS densities are virtually identical. This is expected, since their formulas are mirror images of each other (and since we simulated the number of locations and industries to be the same). Note also that the mean of the two densities is a slightly stronger predictor than either one individually which suggests that there is some information in each measure that is not captured by the other.

**Figure 1: Observed vs. Implied Comparative Advantage for Different Noise Levels**



Up until this point, we have fixed  $k$  (the number of comparators used to build the density variable) equal to  $\sqrt{N}$ . We now relax this constraint, in order to see what happens as  $k$  varies. Figure 1 presents the results, across three levels of noise. As expected, we find an inverted-U shaped relationship between  $k$  and the correlation between *true* and *implied* comparative advantage: the correlation grows stronger as we include more comparators, then quickly peaks around the lower-middle range,  $k = 8$  to  $35$  (depending on the level of noise). After that, adding more comparators gradually weakens the correlation. Reassuringly, we can also note that  $R^2$  values are relatively stable across the intermediate  $k$  range (*i.e.*, the curves are flat in the middle); this suggests that our results will not be sensitive to

<sup>12</sup>We also test the power of our predictors when given pure noise as an input (*i.e.*, without the underlying Ricardian component). As expected, the  $R^2$  values fall to zero.

small changes in  $k$ , as long as we stay in that lower-middle range (including our default  $k = \sqrt{N}$  setting).

## 1.5 Hypotheses

Before beginning our empirical investigation, we can use these outcomes from our theoretical section and simulations to set some hypotheses:

**H1:** The *implied* comparative advantage measures (*i.e.*, PS and CS density values) should be strongly positively correlated with current *observed* comparative advantage, with estimated coefficients near unity.

**H2:** The *implied* comparative advantage measures contain information on the *true* comparative advantage of industry-locations; as such, the gap between *implied* comparative advantage and *observed* comparative advantage (*i.e.*, the regression residuals) will be positively correlated with industry-location growth over long periods.

**H3:** The *implied* comparative advantage measures can be calculated even for the industries currently absent from a location. In these cases, we can expect *implied* comparative advantage to be predictive of which new industries will emerge (or fail to emerge) in a location.

**H4:** As symmetric measures, the PS and CS density variables will have some non-overlapping information, meaning that we expect higher explanatory power from regressions which include them both.

**H5:** When constructing the similarity indices, we can expect them to be symmetrically distributed about the midpoint (0.5, *i.e.*, random), with either two symmetric peaks (indicating a world conforming closely to the true Ricardian structure) or a single peak at the midpoint (indicating a *noisy* world, with relatively large disturbances away from the true Ricardian ideal).

**H6:** When constructing the density variables, we can expect an inverted U-shaped relationship between the number of comparators used ( $k$ ) and the explanatory power of the regressions; that is, using an intermediate number of comparators (close to the geometric midpoint) should result in the strongest relationship between *implied* and *true* comparative advantage.

## 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 first study the growth rates of industry-location cells, which can only be defined for cells that start with a nonzero output. Later, we study the extensive margin by looking at the appearance of industries that were not initially present 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 begin by using trade data to study the industry-location relationship at an international scale. Here we use UN COMTRADE data, cleaned according to the process detailed in Bustos and Yildirim (2019b). Exports are disaggregated into product categories according to the Harmonized System four-digit classification (HS4), for the years 1995-2016.<sup>13</sup> We restrict our sample to countries with population greater than one million and total exports of at least \$1 billion in 2005 (the midpoint of the period studied). We remove Iraq (which was war-torn and has severe quality issues), Serbia-Montenegro (which split into two countries during the period studied), and Namibia and Botswana (which lack customs data for the initial five years of the period). We then drop two products, “Natural cryolite or chiolite” (HS4 code 2527), and “Vegetable materials used for brooms” (HS 1403) as their world trade both fall to zero in the mid-2000s. We also exclude the miscellaneous code HS 9999 (“Commodities not specified according to kind”). These restrictions reduce the sample to 119 countries and 1239 products that account for 92% of world trade and 93% of the world population in 2016.

In addition to the international trade data, we use data from three countries, namely USA, India and Chile, that quantify the presence of industries in subnational locations. For USA, we use the US Census County Business Patterns (CBP) database from 2003-2011.<sup>14</sup> It includes data on employment and number of establishments by county, which we aggregate into 708 commuting zones (CZ; Tolbert and Sizer (1996)), and 1,086 industries (NAICS 6-digit). This dataset also provides annual payroll data for 698 CZ and 941 NAICS industries.<sup>15</sup> Our Chilean dataset comes from the Chilean tax authority, *Servicio*

---

<sup>13</sup>Trade data for earlier years are available, but it uses a different product classification system. We might introduce error due to major continuity breaks when converting between classifications.

<sup>14</sup>During these years, two versions of NAICS (2002 and 2007) were used. If we extend the data either to earlier or later years, then we would need to convert an additional classification which would be a potential error source.

<sup>15</sup>The discrepancy between employment and establishment versus payroll sample sizes comes from the

*de Impuestos Internos*, and includes the number of establishments based on tax residency for 334 municipalities and 681 industries, from 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.<sup>16</sup> For all the datasets above, we include only industries and regions that have non-zero totals for each year (as we do with the international export data). This approach effectively removes discontinued or obsolete categories.

## 2.2 Constructing the model variables

First, we build the similarity and density matrices for the *implied* comparative advantage introduced above for each dataset. Our first step is to normalize the export, employment and payroll data to focus on the intensity of each industry-location link, and to facilitate comparison across location, industry and time. For the international data, 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 using the population of a location as a measure of its size rather than its total production or exports (Bustos et al., 2012). This small change eliminates the impact of the movement in output or prices of one industry on the values of other industries. We formally define the Revealed per-Capita Advantage (RpCA)<sup>17</sup> of location  $l$  in industry  $i$  in year  $t_0$  as:

$$R_{il,t_0} = \frac{y_{il,t_0} / pop_{l,t_0}}{\sum_l y_{il,t_0} / \sum_l pop_{l,t_0}} \quad (2.1)$$

where  $y_{il,t_0}$  is the export, employment or payroll value,  $pop_l$  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 against high-population locations, we cap  $R_{il}$  at  $R_{max} = 5$  when building our similarity indices (Equations 2.2 and 2.3 below).<sup>18</sup> 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

---

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 range’s 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.

<sup>16</sup>This dataset was constructed by Sam Asher and Paul Novosad, who kindly shared it with us.

<sup>17</sup>Our results are robust to the use of standard RCA instead of RpCA.

<sup>18</sup>We specifically set the ceiling at  $R_{max} = 5$  because this is the highest possible RpCA value for the most populous country in the world, China. In a hypothetical industry  $i$  where China exports the entire industry’s output, then  $R_{i,China} = pop_{World} / pop_{China} \approx 5$  on average over the period studied.

similarity indices defined above:

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

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

In words, two industries are similar if different locations tend to have them in similar intensities. 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 co-agglomeration index (Ellison and Glaeser, 1999).

**Table 2: Most similar location pairs, international trade, 2010**

Location $l$		Location $l'$		Location Similarity
COD	Congo, DR	COG	Congo	0.8081
CIV	Côte d'Ivoire	CMR	Cameroon	0.7987
CIV	Côte 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

Tables 2 and 3 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. (2014) that the machinery-related industries are highly interconnected.

Figure 2 shows the full distribution of the similarity values. The right panel gives the distribution of the location similarity values. The distribution is roughly symmetric, with a peak at the center (median = 0.511, mode at 0.49-0.50). This is well aligned with what our theoretical model would predict, given a *noisy* world (*i.e.*, relatively large random deviations from the underlying *true* comparative advantage) – see Figure A.1 in Appendix. However, the same cannot be said for the industry similarity values (left panel). Their

distribution is also roughly symmetric, but distributed around a peak greater than 0.5 (median = 0.751, mode at 0.75-0.76); this violates the predictions of our theory. Why might the industry similarity values have a higher than expected proportion of positive correlations? One possibility is that the distribution of the underlying industry technology parameters ( $\psi_i$ ) are not in fact random; indeed, the literature describes a *nested* structure of industry cooccurrence (Bustos et al., 2012). Incorporating such a structure may be a promising area for future research (though out of the scope of this paper).

**Table 3: 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

Moving forward, we are particularly interested in the best comparators for each industry or location. Figure 2 illustrates the distribution of each industry or location's most similar comparator, as well as its  $k$  most similar industries or locations (where  $k$  is set to the square root of the total number of industries or locations). In the left panel, we see that each industry's most similar comparator tends to be quite highly correlated with it. Values are similarly high when we extend the scope to each industry's 35 most similar comparators (interquartile range (IQR) = 0.82 to 0.92). In the right panel, we see that the most similar comparator locations tend to have lower similarity values. This includes the top 11 comparators (IQR = 0.59 to 0.64). This suggests that our CS density measure may underperform relative to the PS density measure. More generally, we can see much improvement to our set of comparators by limiting the scope to the most similar subset of industries and locations. This illustrates our motivation for including the nearest neighbor filters to remove poor comparators from consideration.

Having built our similarity indices, we can use them to recreate our density indices

from Equations 1.9 and 1.11, replacing the  $r_{il,t_0}$  with  $R_{il,t_0}$ :

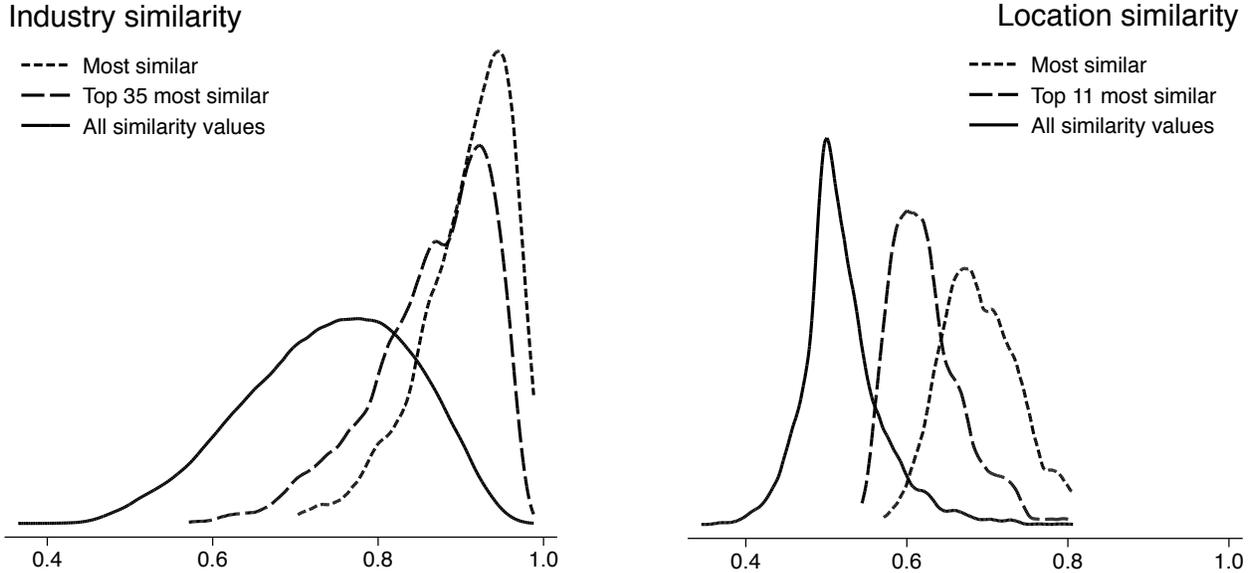
$$w(u)_{il}^{[PS]} = \sum_{i' \in I_{iu}} \frac{\phi_{ii'}}{\sum_{i'' \in I_{iu}} \phi_{ii''}} R_{i'l,t_0} \quad (2.4)$$

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

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

with the set  $L_{lv}$  is the  $v$  nearest neighbors of location  $l$ . As before, we set the neighborhood size  $u$  for product space to  $\sqrt{N_i} \approx 35$  and  $v$  to  $\sqrt{N_l} \approx 11$  nearest neighbors following Duda et al. (2012). These indices serve as our proxies for a location's *implied* comparative advantage in an industry.

**Figure 2: Distribution of similarity values.**



### 3 Main empirical results

#### 3.1 Estimating *implied* comparative advantage

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:

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

where  $\varepsilon_{il,t_0}$  is the residual term.

**Table 4: OLS regression of international exports by industry-location, 1995**

	(1)	(2)	(3)
	Observed Comparative Advantage RpCA of Exports (log), 1995		
Product Space density, 1995 (log)	0.972*** (0.013)		0.785*** (0.018)
Country Space density, 1995 (log)		0.988*** (0.043)	0.285*** (0.021)
Observations	94,702	94,702	94,702
Adjusted R-squared	0.635	0.481	0.651

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

Table 4 shows that both the PS and CS density terms are highly significant ( $p < 0.001$ ), with coefficients very close to unity. As expected, the terms also explain a very large fraction of the variance of the country-product export intensity, though the PS density generates a significantly higher  $R^2$  than the CS density. When included in regressions together, both terms are still highly significant (indicating some non-overlapping information, as expected), and they explain nearly two thirds of the variation in export intensity.

Table 5 shows the regressions for the US, India and Chile datasets. In all cases, both PS and CS density are significant ( $p < 0.01$  in all cases), again signaling that they contain non-overlapping information; interestingly, they also yield coefficients that sum close to unity on average (from 0.81 to 1.29). As with the export data,  $R^2$  values are substantial, especially for the establishment count datasets. These results validate our first hypothesis: the value of an industry-location's current comparative advantage in our data can be estimated with some accuracy based on the values of similar industries and locations – *i.e.*, through its *implied* comparative advantage. 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 5: 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
	RpCA (log)			log	log
Product Space density (log)	0.620*** (0.010)	0.476*** (0.016)	0.530*** (0.018)	0.293*** (0.021)	0.306*** (0.015)
Country Space density (log)	0.556*** (0.010)	0.363*** (0.018)	0.759*** (0.013)	0.517*** (0.018)	0.531*** (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 \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

### 3.2 Convergence between implied and observed comparative advantage

We have shown that our *implied* comparative advantage measures can explain much of the variation in the world’s current comparative advantage. However, our theoretical model also implies that the residual from this regression – *i.e.*, the gap between *observed* and *implied* comparative advantage – should be 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.<sup>19</sup> We construct the variables as follows. We use the standard definition of the annualized growth rate of  $y_{il}$ :

$$\dot{y}_{il} = \frac{1}{t_1 - t_0} \log \left( y_{il,t_1} / y_{il,t_0} \right) \quad (3.2)$$

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.*,  $y_{il,t_1} = 0$ ) 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  $y_{il,t_0} \neq 0$  and  $y_{il,t_1} \neq 0$ . In Section 3.3, we use a *probit* regression model to examine the probability of industry appearance (*i.e.*, growth from zero) and

<sup>19</sup>Our results improve somewhat if we include the *implied* comparative advantage value here instead of the residual. But our goal here is to show that even the residual term is predictive of the future growth.

disappearance (*i.e.*, collapse into zero).

Our growth regression takes the following form, based on Equation 1.13:

$$\dot{y}_{il} = \alpha + \beta_{\varepsilon}\varepsilon_{il,t_0} + \gamma c_l + \delta d_i + e_{il}. \quad (3.3)$$

Here  $\varepsilon_{il,t_0}$  is the residual term of the regression from the first stage, and  $\beta_{\varepsilon}$  is its regression coefficient. Note that while Equation 1.13 estimates the effect of *implied* comparative advantage minus current comparative advantage, our residual term here is the opposite (current comparative advantage minus a function of *implied* comparative advantage); we thus expect  $\beta_{\varepsilon}$  to take a negative sign. The intuition is that if the current comparative advantage is above (below) what we would expect given the *implied* comparative advantage, then we anticipate that the future trend in *observed* comparative advantage will be negative (positive). We have also added a constant term ( $\alpha$ ), and location and industry control variables (with coefficients  $\gamma$  and  $\delta$ , respectively). Finally,  $e_{il}$  is the error term of the regression.

Table 6 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 6 use as independent variable the error terms from the three regressions in Table 4. 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 18 percent of the variance of growth between 1995 and 2016. The residual terms for PS and CS density explain equal proportions of the variance; yet as before, the highest  $R^2$  value comes from both terms together, suggesting that their residuals also contain non-overlapping information. Both terms have the expected negative sign, and are significant at  $p < 0.001$ .

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; these correspond with the industry-level and location-level components of the decomposition in Equation 1.1. Note that these variables are constructed using information from the base year of the regression alone. Column 4 shows that these variables, on their own, are significantly related to subsequent growth, as noted by Glaeser et al. (1992); however, Column 5 indicates that they do not substantially affect the magnitude and significance of the density residuals, and instead see their own significance decrease.

**Table 6: OLS regression of export growth of an industry in a country (1995-2016)**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Growth in exports (log), 1995-2016								
Residual, Product	-0.023***		-0.013***		-0.013***		-0.019***		-0.018***
Space density, 1995	(0.001)		(0.002)		(0.002)		(0.002)		(0.001)
Residual, Country		-0.019***	-0.011***		-0.008***		-0.005***		-0.008***
Space density, 1995		(0.001)	(0.002)		(0.002)		(0.001)		(0.001)
Industry-location exports, 1995 (log)				-0.012***	-0.003**	-0.010***	0.000	-0.021***	-0.000
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Location population 1995 (log)				0.012***	0.005*	0.009***	0.001		
				(0.003)	(0.003)	(0.002)	(0.002)		
Global industry total 1995 (log)				0.014***	0.006***	0.010***	0.002		
				(0.001)	(0.001)	(0.001)	(0.001)		
Radial industry growth 1995-2016 (log)						0.926***	0.970***		
						(0.019)	(0.017)		
Radial location growth 1995-2016 (log)						0.773***	0.967***		
						(0.109)	(0.115)		
Industry FE								Yes	Yes
Location FE								Yes	Yes
Observations	94,702	94,702	94,702	94,702	94,702	94,702	94,702	94,698	94,698
Adjusted R-squared	0.149	0.149	0.175	0.121	0.182	0.240	0.330	0.412	0.440

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

Next, we introduce controls that account for information beyond the base year, namely the overall rate of growth for each location and each industry; we refer to these controls as the *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 industry's total (summed across all locations):

$$\dot{b}_i = \frac{1}{t_1 - t_0} \log \left( \frac{\sum_l y_{il,t_1}}{\sum_l y_{il,t_0}} \right) \quad (3.4)$$

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:

$$\dot{b}_l = \frac{1}{t_1 - t_0} \log \left( \frac{\sum_i y_{il,t_1}}{\sum_i y_{il,t_0}} \right) \quad (3.5)$$

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 location's exports. Deviations from the radial industry growth mean that countries are changing their global market share in that industry.

We use these radial growth variables for multiple purposes. First, they are an intuitive benchmark comparator for our density indices, as they represent an alternative theory of growth dynamics (balanced growth). 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. Finally, we note that Equation 1.1, when using the RCA functional form, implies that

$$\dot{y}_{il} = \dot{r}_{il} + \dot{b}_i + \dot{b}_l - \dot{W}$$

where as before the dot operator denotes the changes in each logged variable between initial and final period. Specifically,  $\dot{r}_{il}$  is the change in RpCA; this would mean we are then evaluating our residual densities' ability to predict the change in output intensity over time. Nevertheless, it is important to note that benchmarking the performance of the density residuals against radial growth is not a fair comparison, since the density variables are calculated with only base year data to explain growth, while the radial growth

variables use information from the final period as well.

Column 6 shows the effect of radial growth and initial size variables on subsequent growth. 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 over nine percentage points.

In addition to the radial growth variables, we also test our model using industry and location fixed effects. These capture all industry and location specific dynamics, subsuming the size and radial growth control variables as well as any other source of variation at the purely location level or 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 the baseline growth equation with both location and industry fixed effects as well as the initial location-industry size. Column 9 reintroduces the density variables and shows that their economic and statistical significance is undiminished. It is important to again point out that the product and country space residuals can be calculated using only base year data and thus contain no information regarding future growth, while the coefficients on the fixed effects can only be calculated *ex post*. This means that the residuals of the first-stage density regressions still carry new information related to industry-location growth in the subsequent 21 years, even after controlling for all possible industry and location effects.

Finally, it should also be pointed out that 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 7). We find that the PS and CS density residuals are highly significant predictors of industry-location growth, both before and after controlling for initial output, industry and location size, and radial growth ( $p < 0.01$  for all cases). Put together, this seems to suggest that our theoretical model – connecting *implied* comparative advantage to industry-location-level growth – is supported by a variety of datasets.

**Table 7: 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	-0.018***	-0.027***	-0.007***	-0.002***	-0.012***	-0.045***	-0.027***	-0.031***	-0.209***	-0.264***
Space density	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.008)	(0.009)
Residual, Country	-0.026***	-0.025***	-0.022***	-0.024***	-0.034***	-0.024***	-0.009***	-0.020***	-0.104***	-0.064***
Space density	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.009)	(0.007)
Initial industry		0.008***		-0.003***		0.016***		0.007		-0.013
-location level (log)		(0.001)		(0.000)		(0.001)		(0.005)		(0.008)
Initial location		-0.014***		0.003***		-0.026***		0.014***		-0.017
population (log)		(0.001)		(0.000)		(0.002)		(0.002)		(0.012)
Initial global		-0.009***		-0.000		-0.023***		0.023***		0.039***
industry total (log)		(0.001)		(0.000)		(0.002)		(0.001)		(0.009)
Radial industry		0.383***		0.331***		0.379***		0.681***		1.047***
growth (log)		(0.004)		(0.004)		(0.004)		(0.013)		(0.010)
Radial location		0.324***		0.283***		0.210***		0.423***		0.519***
growth (log)		(0.013)		(0.009)		(0.012)		(0.042)		(0.079)
Observations	279,439		278,946		89,378		50,373		49,651	
Adjusted R <sup>2</sup>	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 \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

### 3.3 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. In particular, we expect that our measures of *implied* comparative advantage contain information on the overall fit of an industry in a location, according to the *true* underlying characteristics.

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  $M_{il}$ :

$$M_{il,t_0} = \begin{cases} 1 & y_{il,t_0} \geq 1 \\ 0 & y_{il,t_0} = 0 \end{cases} \quad (3.6)$$

where, as before,  $y_{il,t_0}$  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  $M_{il,t_0} = 1 \rightarrow M_{il,t_1} = 0$ .

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). 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-location pairs as present in 2003, or 42% of the total sample of industry-location cells. Of these present industry-

location pairs, 45,108 became absent by 2011, yielding a disappearance rate of 14%. Likewise, 37,681 industry-location pairs that were absent in 2003 became present by 2011, resulting in an appearance rate of 8.5%. In Chile, 55,347 industry-location pairs were present in 2005, or 24% of the sample. By 2008, 4,762 of these industry-location cells became absent (a disappearance rate of 8.6%) while 11,496 initially absent industry-location cells became present (an appearance rate of 6.7%). Internationally, 46,255 industry-location pairs were present in our base year of 1995, or 31% of the sample. By 2016, 3,984 of these present industry-location pairs became absent (a disappearance rate of 8.6%) while 7,501 of the 83,342 initially absent industry-location pairs became present (an appearance rate of 9.0%).

We can now use our density indices for the *implied* comparative advantage to explain the appearance and disappearance of industries by location. First, we 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:

$$P\left(M_{il,t_0} = 1\right) = \Phi\left(\alpha + \beta_{PS}w_{il}^{[PS]} + \beta_{CS}w_{il}^{[CS]}\right) \quad (3.7)$$

where  $\Phi$  is a normal cumulative distribution function. As in the intensive margin, the model in Equation 3.7 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  $M_{il,t_0}$ :

$$M_{il,t_0} = \widehat{M}_{il,t_0} + \varepsilon_{il,t_0} \quad (3.8)$$

where  $\widehat{M}_{il,t_0}$  is the expected probability of industry presence and  $\varepsilon_{il,t_0}$  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.

**Table 8: 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)	1.286*** (0.035)		1.005*** (0.033)
Country Space density, initial year		0.795*** (0.004)	0.772*** (0.005)		0.939*** (0.005)	0.822*** (0.006)		1.186*** (0.057)	0.416*** (0.031)
All industry-locations		768,888			227,454			147,441	
Present industries		324,622			55,347			46,255	
Presence rate		42.22%			24.33%			31.37%	
Area Under the Curve	0.924	0.940	0.940	0.815	0.900	0.911	0.930	0.894	0.934
Pseudo $R^2$	0.341	0.493	0.495	0.357	0.193	0.454	0.498	0.386	0.516

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

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  $\varepsilon_{il,t_0}$  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 8 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; AUC values are also very high (AUC between 91% and 94% for hybrid models). Likewise, the 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.

Next, we use the residual term from these regressions to predict industry appearances and disappearances over long-term horizons in each dataset (Table 9). 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 model's AUC is 80% for appearances and 72% for disappearances. For the international trade data, over a 21-year horizon, the AUC is 77.3% for appearances and 75.4% for disappearances. Put together, this suggests that the *unexpectedly absent* industries tend to preferentially appear over time while the *unexpectedly present* industries tend to disappear. This supports the hypothesis that our *implied* comparative advantage measures contain information on the *true* Ricardian pattern of presences and absences, based on the hidden underlying technological requirements and endowments.

**Table 9: 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, 1995-2016		
Residual, Product Space density	-2.858*** (0.026)			-2.636*** (0.037)			-0.909*** (0.090)		
Residual, Country Space density		-3.004*** (0.017)			-1.757*** (0.038)			-0.829*** (0.052)	
Residual, hybrid density			-2.994*** (0.017)			-2.389*** (0.031)			-0.973*** (0.083)
Initially absent Industry appearances		444,266 37,681			172,107 11,496			83,342 7,501	
Appearance rate		8.48%			6.68%			9.00%	
Area under the curve	0.801	0.832	0.834	0.757	0.747	0.803	0.7501	0.733	0.773
Pseudo $R^2$	0.059	0.145	0.144	0.064	0.021	0.073	0.0889	0.0763	0.101
	(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, 1995-2010		
Residual, Product Space density	2.953*** (0.018)			0.929*** (0.039)			0.704*** (0.046)		
Residual, Country Space density		2.265*** (0.009)			1.435*** (0.038)			0.727*** (0.064)	
Residual, hybrid density			2.272*** (0.009)			1.368*** (0.030)			0.757*** (0.046)
Initially present Industry disappearances		324,622 45,108			55,347 4,762			46,255 3,984	
Disappearance rate		13.90%			8.60%			8.6%	
Area under the curve	0.840	0.854	0.855	0.625	0.708	0.722	0.742	0.737	0.754
Pseudo $R^2$	0.231	0.249	0.250	0.022	0.066	0.081	0.115	0.101	0.129

Location-clustered robust standard errors in parentheses.

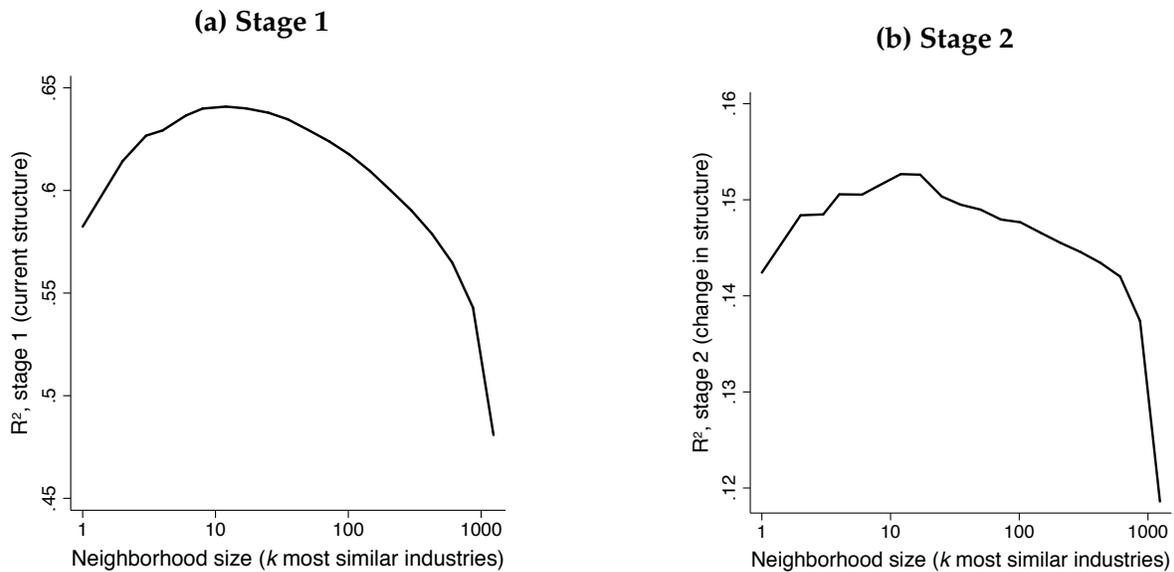
Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## 4 Testing Other Implications of the Model

### 4.1 Number of comparators used

In the international trade data regressions above, we constructed the PS and CS density measures using the top 35 most similar industries ( $k = \sqrt{N_i}$ ) and the top 11 most similar locations ( $k = \sqrt{N_l}$ ), respectively. We can now examine the impact of letting  $k$  vary. In the theory and simulation sections, we set the expectation that we would find an inverted U-shaped relationship between  $k$  and the explanatory power of density; we can now test that hypothesis using real data.

Figure 3: Changing Number of Comparators.



Figures 3a and 3b show the results, looking at how the number of comparator industries used to construct PS density affects its explanatory power, in stage one and stage two regressions, respectively. In both cases, we find a relationship quite similar to our simulations (Figure 1 – see especially the *high noise* results): the  $R^2$  values peak near the geometric midpoint (at  $k=35$  industries), then decrease as more comparator industries are added. We also found a similar pattern for CS density (not shown), with  $R^2$  peaking near  $k = \sqrt{N_l}$ . Thus, we can confirm that our density measures perform as expected based on the theoretical model. Finally, we note that the  $R^2$  values are fairly consistent in the (logarithmic) middle range of  $k$  values (as expected); this verifies that our results are robust to the choice of  $k$  value (within a wide intermediate range).

## 4.2 Using the least similar comparators

Up until this point, we have built the density functions using the most similar comparators for each industry or location. Yet our model would imply that we could also obtain information from the least similar comparators as well.<sup>20</sup> The intuition is that for every comparator who acts similarly to a location or to an industry, there should also be comparators who do the opposite; thus, the *implied* comparative advantage would be the opposite of what the least similar comparators imply.<sup>21</sup> Then as before, the difference between *observed* and *implied* comparative advantage should be negatively associated with growth.

However, looking at the similarity distributions (Figure 2), we would expect this relationship to hold better for the country space density than for product space density. This is because the industry similarities are distributed about 0.75 instead of 0.5 (random), so we would expect that the many of the least similar industries will still be somewhat similar (though much closer to random than the most similar comparators).<sup>22</sup> This would suggest that a density built with the least similar locations would be negatively correlated with current comparative advantage (as expected), while a density built with the least similar industries would have a positive but relatively weak correlation with current comparative advantage.

Here, we create the density variables as before, but using the  $\sqrt{N}$  least similar locations and industries. In cross-sectional regressions (Table 10), we observe that the least similar CS density performs as expected: its sign is negative, and comparable in magnitude to that of the traditional CS density (though with a lower  $R^2$ ). On the industry side, the least similar PS density has a positive sign, though with a relatively smaller magnitude (compared to original PS density), and we can reject the hypothesis that these coefficients are equal to each other with  $p < 0.001$ . Interestingly, when controlling for original PS density, the least similar density switches to a negative coefficient, and increases in significance. We can perhaps imagine that least similar density variable as containing two components: new information coming from genuinely dissimilar industries, and redundant information from weakly similar industries. When combined with a density built from the most similar industries, the redundant information is cancelled out, leaving only the dissimilarity information. This would explain why the coefficients of most similar and least similar industries in Column 6 is consistent with the coefficients

---

<sup>20</sup>Thanks to two anonymous referees for suggesting this test.

<sup>21</sup>For example, the lowest similarity value we observe is between China and the US ( $\phi_{China,USA} = 0.341$ ); this is far below random (0.5), meaning that we can expect one country to export what the other does not.

<sup>22</sup>Specifically, the set of any industry's 35 least similar comparators has an IQR between 0.51 and 0.57 – still slightly higher than random.

of most and least similar locations Column 3.

**Table 10: Regressions with Least Similar Industries.**

	(1)	(2)	(3)	(4)	(5)	(6)
	Observed Comparative Advantage RpCA of Exports (log), 1995					
CS Density - most similar locations, 1995 (log)	0.988*** (0.043)		0.778*** (0.039)			
CS Density - least similar locations, 1995 (log)		-0.879*** (0.045)	-0.352*** (0.034)			
PS Density - most similar industries, 1995 (log)				0.972*** (0.013)		1.005*** (0.015)
PS Density - least similar industries, 1995 (log)					0.221** (0.109)	-0.203*** (0.023)
Observations	94,664	94,664	94,664	94,664	94,664	94,664
Adjusted R-squared	0.481	0.336	0.513	0.635	0.013	0.645

Country-clustered robust standard errors in parentheses.

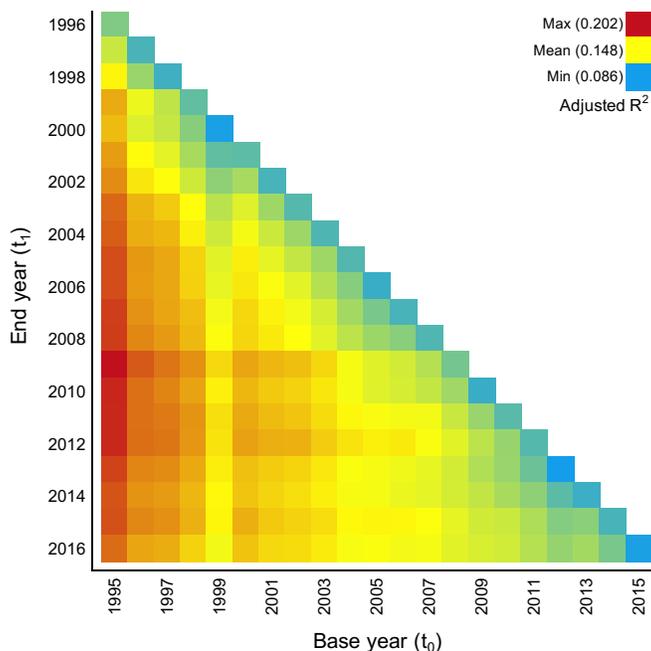
Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

Finally, we can use the residuals from our cross-sectional regressions, which capture the gap between *observed* and *implied* comparative advantage, to explain industry-location growth. In Table A.1, the residuals for the least similar densities perform as expected, for both CS and PS (see table in appendix): they all have negative signs, are comparable in magnitude to the original densities, and are highly significant ( $p < 0.001$  in all cases).

### 4.3 Effect of time horizon

According to our theoretical model, the *implied* comparative advantage variable should be a better predictor of trends in the *observed* comparative advantage in relatively longer time-frames. In our growth regression models above, we chose the base year to be the initial year of the available data and the end year as the final year of the available data. We test the sensitivity of our results reported in Table 6 depending on the choice of different base years and end years. Figure 4 shows the adjusted  $R^2$  values for our international trade regressions (including the base-year controls) over all possible year combinations. Each regression explains a sizable portion of the variation in export growth, with the low-

est adjusted  $R^2$  exceeding 8.6%, and a mean  $R^2$  of 14.8%. Interestingly, we find that predictive power appears to generally improve as the prediction interval increases (barring a possible continuity break between 1999 and 2000). This indicates that the density indices do not capture a short-term mean reversion effect, but a longer-term shift in economic structure. This makes sense looking back at the theoretical model: we were interested in exploring how economies' present-day industrial structures might deviate from their hidden comparative advantage, and whether such deviations would persist in the longer time horizons.



**Figure 4:** Heat map of out of sample predictions of export growth, hybrid density model.

## 5 Evaluating alternative explanations and robustness of the results

### 5.1 Testing the effect of the classification system

Could our results be an artifact of how industries are defined? We note the possibility that our results could also be explained by industry classification: if a classification system arbitrarily splits a single economic activity into two industry categories, then we would expect to see them in similar intensities. In the classification of the international trade data, for example, we can see that HS6101 contains “Men’s overcoats” whereas HS6102

contains “Women’s overcoats.” In fact, in Table 3, 7 out of 10 product pairs have the same 2-digit Harmonized System codes. To explore the possibility that our results are driven by such trivial cases, we calculate an adjusted product space density that excludes industries from same 2-digit Harmonized System categories. We then perform the same regressions as before which lets us zero in on the power of comparators that appear similar (in their export structure), but are not in the same HS category.

Table 11 shows the result for the initial stage. As the table shows, the adjusted density is still highly significant,  $R^2$  values decrease only slightly (from 63.4% to 61.5%), and with the coefficients of the standard and adjusted density terms are statistically indistinguishable (at the 5% level).

We also find little to no effect of this adjustment in the second stage: when we use the residuals from the first stage to predicting export growth between 1995 and 2016, excluding the same 2-digit industries from density does not hurt the predictive power. Comparing Columns 1 to 3 and 2 to 4 of Table A.3, the coefficients of the residual terms are not statistically different from each other at 5% level. Moreover, the explanatory power, which is captured by the  $R^2$  term, is not affected by excluding the same 2-digit products (decreases under 0.02). These results thus appear to reject the possibility that our findings are an artifact of the industry classification used.

For balance, we can also test the explanatory power of taking the mean RpCA values for other products in the same 2-digit category (within the same location). Such a metric would essentially be capturing the information from classification system itself, since it would rely on the system’s designation of which products are similar enough to have the same first two digits. (Our measure, by comparison, takes no information from the classification system.) We expect it to perform well, since it is quite reasonable to estimate the comparative advantage of a location in a product based on that location’s exports of other products in the same category. Column 3 of Table 11 shows the results for the first stage where we predict the current levels and, indeed, the coefficients is highly similar to the PS density, albeit statistically lower. Explanatory power is also lower as manifested by the  $R^2$  values. Comparing Columns 1 to 5 and 2 to 6 of Table A.3, we observe that the PS density is better in explaining the future levels.

Our results indicate that the industry similarity measure based solely on the classification system also has predictive power, but even when we exclude the products from same 2-digit category, our predictive power remains high. Hence, our PS density captures more than the classification system indicates.

**Table 11: Building densities from other industry similarity matrices.**

	(1)	(2)	(3)	(4)	(5)	(6)
	Observed Comparative Advantage RpCA of Exports (log), 1995					
Product Space density 1995 (log)	0.972*** (0.013)					
PS density – excluding same 2-digit industry, 1995 (log)		0.953*** (0.014)				
Mean of other R.I in same 2-digit industry, 1995 (log)			0.935*** (0.006)			
Taxonomy (using R.I) 1995 (log)				0.711*** (0.016)		
Taxonomy (Using RCA) 1995 (log)					0.620*** (0.030)	
PS Density, HK method 1995 (log)						2.303*** (0.185)
Observations	94,765	94,765	94,221	51,079	51,079	94,765
Adjusted R-squared	0.634	0.615	0.607	0.496	0.218	0.367

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## 5.2 Using other similarity matrices to build density.

We have thus far shown evidence that the intensity of an industry in a location can be proxied using the output of highly similar industries in that same location – its *implied* comparative advantage. We are somewhat agnostic to the exact functional form for our proxy: we do not claim that our similarity matrices or density formula are more accurate than other measures, but rather that it can be derived from a Ricardian-inspired model of the world. Here, we show that the patterns predicted by our theoretical model should also apply to other density measures.

Specifically, we will test the product space density proposed by Hausmann and Klinger (2006) and the taxonomy measure proposed by Zaccaria et al. (2014). Original product space density uses the binary presence-absence matrix to calculate similarity measure based on co-production probabilities. The density variable is calculated using all information in the matrix. The taxonomy matrix, on the other hand, starts from the same

presence-absence matrix but uses a different normalization factor. Then, from the similarity matrix, it selects the top-entry to define the most important contributor to an industry.

We again expect the density measures to perform similarly to our measure; however, their density is built using the binary presence-absence matrix, rather than with continuous RCA values; this may result in a loss of explanatory power. We will also no longer expect the coefficient of HK Product Space to approach unity, since its expected scale is not the same as the dependent variable; the coefficient should still be positive and highly significant, however. Table 11 shows the results for the initial stage. As expected, all terms are positive and highly significant. We also find little to no effect of these adjustments in the second stage: Table A.3 shows that when we use the residuals from the first stage to predicting export growth between 1995 and 2016, the residual term is always negative (as expected), highly significant, and with a similar magnitude to that of our original density formula. These results thus appear to reject the possibility that our findings are an artifact of the industry classification used, or the exact density formula.

### 5.3 Effect of the diversity of countries – Yule process

Thus far, we have provided evidence that our measure of *implied* comparative advantage is associated with both current and future comparative advantage levels as well as industry appearances and disappearances. But does this explanatory power comes from our patterns of industry and location similarity, or is another underlying force active? Specifically, here, we will test whether a Yule-like process is driving our results. In ecology, the Yule process is a birth process where each individual organism has the same probability of giving a birth, resulting in organisms with larger size to dominate the population in the future. In our model, the counterpart would be that each industry in a location has the same probability of initiating a new industry; therefore, locations with higher number of present industries (higher diversity) today might be likelier to see appearances in the future.<sup>23</sup> Likewise, we would expect the number of disappearances (deaths in ecology) to decrease with the diversity.

To test this, we include the diversity variable (number of *present* industries in a location) in our probit regressions where we predict the industry appearances and disappearances as in Table 9. Table 12 shows that the diversity is a positive predictor of the appearance but a negative predictor of disappearances, as expected in a Yule world. When put together with the hybrid density, both variables remain significant with the expected signs. Our results imply that a Yule-like process may indeed exist, but cannot

---

<sup>23</sup>Thanks to an anonymous referees for the idea.

solely explain the component of industry-location growth linked with our measures.

**Table 12: Controlling for Yule Process: Probit regression of industry-location extensive margin with diversity measure**

	(1)	(2)	(3)	(4)	(5)	(6)
	International (exports), 1995-2016					
	Industry appearances			Industry disappearances		
Residual, Hybrid density	-0.973***		-0.569***	0.757***		0.531***
1995	(0.083)		(0.117)	(0.046)		(0.051)
Number of industries present		0.899***	0.509***		-1.167***	-0.492***
1995 (log)		(0.140)	(0.181)		(0.076)	(0.098)
Observations	81,665	81,665	81,665	46,232	46,232	46,232
Area under the curve	0.773	0.735	0.760	0.754	0.755	0.764
Pseudo R2	0.101	0.100	0.116	0.129	0.111	0.136

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## 5.4 Cross validation

In order to calculate our density variables (measuring *implied* comparative advantage), we exclude the location or industry being proxied from the weighted average. However, other information regarding that location or that industry is also used in the calculations of the similarity matrices. 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 machine learning literature. In this approach, we build the density indices using only information found in the training set. 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 the testing data.

Table 13 applies this process to our international trade dataset over the 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). Further-

more, 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 13: Cross-validating OLS regression of international exports and export growth by industry-location.**

	(1)	(2)	(3)	(4)	(5)	(6)
	Observed Comparative Advantage RpCA of Exports (log), 1995			Growth in exports log, 1995-2016		
Product Space density out-of-sample, 1995 (log)	0.916*** (0.025)		0.940*** (0.065)			
Country Space density out-of-sample, 1995 (log)	0.150*** (0.038)		0.063 (0.046)			
Product Space density in-sample, 1995 (log)		0.830*** (0.029)	-0.035 (0.066)			
Country Space density in-sample, 1995 (log)		0.357*** (0.049)	0.121** (0.053)			
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)
Observations	23,794	23,794	23,794	23,794	23,794	23,794
Adjusted R-squared	0.622	0.557	0.622	0.187	0.185	0.189

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## 5.5 Product and Country Groups

We can also ask if this theoretical framework is more powerful within certain subsets of industries or locations. We begin by calculating the similarities and densities as before, using the full set of the 1995 export data. Next, we calculate the stage one and two regres-

sions as before (and with the base-year and radial growth controls), but restricted only to the subsamples. We then report the resulting adjusted  $R^2$ . That is, our results for the wood products subsample measures how well our standard density variables can predict export growth for goods in the wood products category alone.

**Table 14: Cross-sectional Regressions for Product and Country Groups.**

Products		Countries	
HS Chapter	Adjusted $R^2$	By income	Adjusted $R^2$
Electronics, Machinery & Equipment	0.449	Low income	0.379
Medical, Consumer & Other	0.435	Upper middle income	0.355
Plastics & Rubbers	0.393	High income	0.327
Processed Metals	0.326	Lower middle income	0.296
Processed Stone & Glass	0.323	By region	Adjusted $R^2$
Chemicals & Related	0.319	North America	0.427
Apparel & Textiles	0.302	East Asia & Pacific	0.414
Wood Products	0.299	Europe & Central Asia	0.398
Automotive, Planes, Ships & Related	0.292	South Asia	0.308
Processed Foodstuffs	0.285	Middle East & North Africa	0.266
Agricultural Products	0.251	Latin America & Caribbean	0.259
Extractives	0.202	Sub-Saharan Africa	0.255

Table 14 shows the results. For product subsamples (based on the HS chapters), it appears that the most easily-explained categories are high-tech goods like electronics or medical devices. The worst predictions are in the extractives and agricultural categories; this makes sense, since shifts in these commodities may have more to do with geographic luck (*e.g.*, oilfield discoveries) than shared technological requirements.

Next, we can divide countries by income level (according to World Bank Group classifications): the groups are relatively close to each other, though low-income countries are the most predictable under our framework (possibly because they are less likely to shift their comparative advantage over the period). Finally, the results by region appear to fall roughly in order of (non-oil) income (unlike looking at income directly); this would make Latin America and the Caribbean somewhat less predictable than expected based on income alone.

## 5.6 Using RCA instead of RpCA.

We chose to use RpCA rather than more traditional RCA measure because RCA could be affected by the price movements in other industries. To be explicit, suppose that a country is a successful exporter of a product in year  $t_0$  with  $RCA > 1$ . And suppose the country's share in a product remains constant in a later year, but because of the movements in the commodity prices, the country's total exports increase as a share of the world exports. Although nothing fundamental changes in the country's ability to make this product because of the price movements in other products the country's RCA becomes smaller.

As a robustness check, we developed all our measures with Balassa's RCA as well. Table 15 shows that both Product Space and Country Space based on RCA are significant predictors, yet, the explanatory powers of these variables are lower than their counterparts using RpCA in Table 4.

**Table 15: OLS regression of international exports by industry-location using RCA based density, 1995**

	(1)	(2)	(3)
	Observed Comparative Advantage RpCA of Exports (log), 1995		
Product Space Density (log), 1995	0.967*** (0.013)		0.811*** (0.019)
Country Space Density (log), 1995		0.726*** (0.038)	0.232*** (0.021)
Observations	93,984	93,984	93,984
Adjusted $R^2$	0.416	0.258	0.432

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

When we do the growth regressions as reported in Table 6, the residual in the first stage performs equivalently with its RcPA counterpart (Table A.2).

## 6 Conclusions

In this paper we have presented and tested a model showing that the *observed* comparative advantage of an industry in a location 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, we present evidence backing up our model's prediction that the error term in the predicted pattern is not pure noise, but instead should carry information regarding the future trend of that industry-location cell. These dynamics include components that are orthogonal to pure industry or location effects, but instead capture industry-location interactions. These results can be found using international trade data as well as sub-national data for the USA, India and Chile. We have shown evidence 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 most intensely at long horizons of over a decade.

Our methodology combines the methods from international trade and relatedness-based measures often used in the regional development. We began by building a modified Ricardian model in which the industry comparative advantage 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 host the industry.

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. The interpretation we model 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 (Glaeser et al., 1992; Ellison et al., 2010; Beaudry and Schiffauerova, 2009). 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.

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 may 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 diminish over time as firms exploit technological spillovers.

Ricardian models are reduced-form models, where other elements are subsumed in the labor productivity parameters. In the Appendix, we show that we can motivate our approach also with a model with an indeterminate number of factors of production. From a factor based model 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.

The results in our paper may be important for policymakers making decisions related to industrial growth strategy. First, it gives the policymakers a tool to estimate the potential comparative advantage level even the industry is not present in the location. Second, it will also help identify industries over-performing or under-performing depending on the levels of other industries present in the location. Third, it may help the policymakers to estimate the growth potential of industries. Fourth, it might also help policymakers to identify pervasive failures and obstacles using historical data to identify which industries had the potential to grow and why the location has not achieved the potential in that industry. Then, these failures or obstacles could be addressed using policy tools.

## References

- Antweiler, Werner and Daniel Trefler**, "Increasing Returns and All That: A View from Trade," *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," *Journal of Political Economy*, 1964, 72 (6), 584–596.
- Beaudry, Catherine and Andrea Schifffauerova**, "Who's right, Marshall or Jacobs? The localization versus urbanization debate," *Research Policy*, 2009, 38 (2), 318–337.
- Boschma, Ron**, "Relatedness as driver of regional diversification: A research agenda," *Regional Studies*, 2017, 51 (3), 351–364.
- **and Gianluca Capone**, "Institutions and diversification: Related versus unrelated diversification in a varieties of capitalism framework," *Research Policy*, 2015, 44 (10), 1902–1914.
- , **Asier Minondo, and Mikel Navarro**, "Related variety and regional growth in Spain," *Papers in Regional Science*, 2012, 91 (2), 241–256.
- , – , **and –** , "The Emergence of New Industries at the Regional Level in Spain: A Proximity Approach Based on Product Relatedness," *Economic Geography*, 2013, 89 (1), 29–51.
- , **Gaston Heimeriks, and Pierre-Alexandre Balland**, "Scientific knowledge dynamics and relatedness in biotech cities," *Research Policy*, 2014, 43 (1), 107–114.
- Bowen, Harry P, Edward E Leamer, and Leo Sveikauskas**, "Multicountry, multifactor tests of the factor abundance theory," *American Economic Review*, 1987, 77 (5), 791–809.
- Bustos, Sebastián and Muhammed A Yildirim**, "Production Ability and Economic Growth," Technical Report, Center for International Development at Harvard University Fellows Working Paper No:110 2019.
- **and –** , "Uncovering trade flows," 2019. Unpublished mimeo, available upon request.

- , **Charles Gomez, Ricardo Hausmann, and César A Hidalgo**, “The Dynamics of Nest-  
edness Predicts the Evolution of Industrial Ecosystems,” *PloS one*, 2012, 7 (11), e49393.
- Caldarelli, Guido, Matthieu Cristelli, Andrea Gabrielli, Luciano Pietronero, Antonio  
Scala, and Andrea Tacchella**, “A network analysis of countries? export flows: firm  
grounds for the building blocks of the economy,” *PloS one*, 2012, 7 (10), e47278.
- Caliendo, Lorenzo, Fernando Parro, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte**,  
“The impact of regional and sectoral productivity changes on the US economy,” *The  
Review of Economic Studies*, 2017, 85 (4), 2042–2096.
- Conway, Patrick J**, “The case of the missing trade and other mysteries: Comment,” *Amer-  
ican Economic Review*, 2002, 92 (1), 394–404.
- Costinot, Arnaud, Dave Donaldson, and Cory Smith**, “Evolving comparative advantage  
and the impact of climate change in agricultural markets: Evidence from 1.7 million  
fields around the world,” *Journal of Political Economy*, 2016, 124 (1), 205–248.
- , – , 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.
- Cristelli, Matthieu, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Lu-  
ciano Pietronero**, “Measuring the intangibles: A metrics for the economic complexity  
of countries and products,” *PloS one*, 2013, 8 (8), e70726.
- Davis, Donald R and David E Weinstein**, “An Account of Global Factor Trade,” *American  
Economic Review*, 2001, 91 (5), 1423–1453.
- and **Jonathan I Dingel**, “The comparative advantage of cities,” Technical Report, Na-  
tional Bureau of Economic Research 2014.
- , **David E Weinstein, Scott C Bradford, and Kazushige Shimpo**, “Using International  
and Japanese Regional Data to Determine When the Factor Abundance Theory of Trade  
Works,” *American Economic Review*, 1997, 87 (3), 421–46.
- Deardorff, Alan V**, “The general validity of the Heckscher-Ohlin theorem,” *American Eco-  
nomic Review*, 1982, 72 (4), 683–694.
- , “Testing trade theories and predicting trade flows,” *Handbook of International Eco-  
nomics*, 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 –, "Defining clusters of related industries," *Journal of Economic Geography*, 2015, 16 (1), 1–38.
- Dornbusch, Rudiger, Stanley Fischer, and Paul Anthony Samuelson**, "Comparative advantage, trade, and payments in a Ricardian model with a continuum of goods," *American Economic Review*, 1977, 67 (5), 823–839.
- Duda, Richard O, Peter E Hart, and David G Stork**, *Pattern classification*, John Wiley & Sons, 2012.
- 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?," *American Economic Review*, 1999, 89 (2), 311–316.
- , –, and **William R Kerr**, "What Causes Industry Agglomeration? Evidence from Co-agglomeration Patterns," *American Economic Review*, 2010, 100 (3), 1195–1213.
- Feenstra, Robert C**, *Advanced international trade: theory and evidence*, Princeton University Press, 2003.
- 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.
- Hanlon, W Walker and Antonio Miscio**, "Agglomeration: A long-run panel data approach," *Journal of Urban Economics*, 2017, 99, 1–14.
- Hausmann, Ricardo and Bailey Klingler**, "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, Alexander Simoes, and Muhammed A. Yıldırım**, *The Atlas of Economic Complexity: Mapping Paths to Prosperity*, The MIT Press, 2014.
- 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 and Ricardo Hausmann**, “The building blocks of economic complexity,” *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106 (26), 10570–10575.
- , **Bailey Klinger, A-L Barabási, and Ricardo Hausmann**, “The product space conditions the development of nations,” *Science*, 2007, 317 (5837), 482–487.
- , **Pierre-Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, Canfei He, Dieter F Kogler, Andrea Morrison et al.**, “The principle of relatedness,” in “International Conference on Complex Systems” Springer 2018, pp. 451–457.
- 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.
- Liang, Jiaochen and Stephan J Goetz**, “Technology intensity and agglomeration economies,” *Research Policy*, 2018, 47 (10), 1990–1995.
- Linden, Greg, Brent Smith, and Jeremy York**, “Amazon. com recommendations: Item-to-item collaborative filtering,” *Internet Computing, IEEE*, 2003, 7 (1), 76–80.
- Lu, Ren, Min Ruan, and Torger Reve**, “Cluster and co-located cluster effects: An empirical study of six Chinese city regions,” *Research Policy*, 2016, 45 (10), 1984–1995.
- Marshall, Alfred**, *Principles of economics*, Macmillan and Company, 1890.

- 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.
- Petralia, Sergio, Pierre-Alexandre Balland, and Andrea Morrison**, "Climbing the ladder of technological development," *Research Policy*, 2017, 46 (5), 956–969.
- Porter, Michael**, "The economic performance of regions," *Regional Studies*, 2003, 37 (6-7), 545–546.
- 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.
- Tacchella, Andrea, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero**, "A new metrics for countries' fitness and products' complexity," *Scientific reports*, 2012, 2, 723.
- , –, –, –, –, and –, "Economic complexity: conceptual grounding of a new metrics for global competitiveness," *Journal of Economic Dynamics and Control*, 2013, 37 (8), 1683–1691.
- 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," *American Economic Review*, 1995, 85 (5), 1029–1046.
- **and Susan Chun Zhu**, "Beyond the algebra of explanation: HOV for the technology age," *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.
- Zaccaria, Andrea, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero**, "How the taxonomy of products drives the economic development of countries," *PloS one*, 2014, 9 (12), e113770.
- 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.

# A Appendix

## A.1 Additional Tables

Table A.1: Growth Regressions with Least Similar Industries

	(1)	(2)	(3)	(4)
	Growth in exports (log), 1995-2010			
Residual, CS density – most similar locations, 1995	-0.019*** (0.001)			
Residual, CS density – least similar locations, 1995		-0.015*** (0.001)		
Residual, PS density – most similar industries, 1995			-0.023*** (0.001)	
Residual, PS density – least similar industries, 1995				-0.012*** (0.001)
Observations	94,664	94,664	94,664	94,664
Adjusted R-squared	0.149	0.117	0.150	0.118

**Table A.2: OLS regression of export growth of an industry in a country using RCA based density**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Growth in exports (log), 1995-2016								
Residual, Product	-0.023***		-0.015***		-0.015***		-0.016***		-0.016***
Space density, 1995	(0.001)		(0.002)		(0.002)		(0.001)		(0.001)
Residual, Country		-0.020***	-0.009***		-0.003*		-0.009***		-0.009***
Space density, 1995		(0.001)	(0.002)		(0.002)		(0.002)		(0.001)
Industry-location exports, 1995 (log)				-0.012***	-0.005***	-0.011***	0.000	-0.021***	-0.003***
Location population 1995 (log)				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Global industry total 1995 (log)				0.012***	0.007**	0.011***	0.004**		
Radial industry growth, 1995-2016 (log)				(0.003)	(0.003)	(0.002)	(0.002)		
Radial location growth, 1995-2016 (log)				0.014***	0.009***	0.011***	0.003**		
				(0.001)	(0.001)	(0.001)	(0.001)		
Industry FE								Yes	Yes
Location FE								Yes	Yes
Observations	93,984	93,984	93,984	93,984	93,984	93,984	93,984	93,984	93,984
Adjusted R <sup>2</sup>	0.146	0.135	0.156	0.121	0.174	0.234	0.320	0.411	0.438

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

**Table A.3: Growth regressions with various density measures.**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Growth in exports (log), 1995-2016											
Residual, PS density 1995	-0.023*** (0.001)	-0.022*** (0.001)										
Industry-location exports 1995 (log)		-0.001 (0.001)		-0.001 (0.001)		-0.002** (0.001)		-0.005*** (0.001)		-0.008*** (0.001)		0.000 (0.002)
Location population 1995 (log)		0.003* (0.002)		0.003 (0.002)		0.003 (0.002)		0.004** (0.002)		0.006*** (0.002)		-0.007*** (0.002)
Global industry total 1995 (log)		0.004*** (0.001)		0.003** (0.001)		0.004*** (0.001)		0.005*** (0.001)		0.008*** (0.001)		0.002 (0.002)
Radial industry growth 1995-2016 (log)		0.978*** (0.017)		0.976*** (0.017)		0.981*** (0.017)		0.976*** (0.021)		0.962*** (0.022)		0.959*** (0.018)
Radial location growth 1995-2016 (log)		0.985*** (0.110)		0.988*** (0.111)		0.927*** (0.106)		0.864*** (0.109)		0.783*** (0.115)		0.645*** (0.115)
Residual, PS density excl. same 2-digit, 1995			-0.022*** (0.001)	-0.022*** (0.001)								
Residual, same 2-digit avg., 1995					-0.022*** (0.001)	-0.020*** (0.001)						
Residual, Taxonomy (using R.I), 1995							-0.017*** (0.001)	-0.011*** (0.001)				
Residual, Taxonomy (Using RCA), 1995									-0.014*** (0.001)	-0.004*** (0.001)		
Residual, PS density (HK method), 1995											-0.018*** (0.001)	-0.017*** (0.002)
Observations	94,765	94,765	94,765	94,765	94,221	94,221	51,079	51,079	51,079	51,079	94,765	94,765
Adjusted R-squared	0.150	0.325	0.148	0.322	0.150	0.316	0.116	0.284	0.121	0.261	0.167	0.287

Country-clustered robust standard errors in parentheses.

Significance given as \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

## A.2 Calculating Expected Similarity Coefficient

### A.2.1 Calculation of expected similarity coefficient for the unit sphere

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:

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

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

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2$$

where  $\text{corr}$  is defined as

$$\text{corr}\{r_i, r_{i'}\} = \frac{\sum_l (r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})}{\sqrt{\sum_l (r_{il} - \bar{r}_i)^2 \sum_l (r_{i'l} - \bar{r}_{i'})^2}}$$

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

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'}) | \psi_i, \psi_{i'}]}{\sqrt{E[(r_{il} - \bar{r}_i)^2 | \psi_i] E[(r_{i'l} - \bar{r}_{i'})^2 | \psi_{i'}]}}$$

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

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'}) | \psi_i, \psi_{i'}]}{E[(r_{il} - \bar{r}_i)^2 | \psi_i]} \quad (\text{A.1})$$

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

$$E[(\tilde{r}_{il} - \bar{\tilde{r}}_i)^2 | \psi_i] = E[\tilde{r}_{il}^2 | \psi_i] - E^2[\tilde{r}_{il} | \psi_i]$$

We can calculate the denominator in Equation A.1 using these separate terms. First,

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

and

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

Hence, the denominator in Equation A.1 becomes:

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

We can write the numerator in Equation A.1 as:

$$\begin{aligned} E[(\tilde{r}_{il} - \bar{r}_i)(\tilde{r}_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}] &= \int_0^1 \left( d^2(\psi_i, \lambda_l) - \frac{1}{12} \right) \left( d^2(\psi_{i'}, \lambda_l) - \frac{1}{12} \right) d\lambda_l \\ &= \int_0^1 [d(\psi_i, \lambda_l)d(\psi_{i'}, \lambda_l)]^2 d\lambda_l - \frac{1}{144} \end{aligned} \quad (\text{A.2})$$

To calculate the integral, we will measure all the distances on the circle relative to  $\psi_i$ . Let's define  $\Delta_{ii'} \equiv d(\psi_i, \psi_{i'})$ . We can write the integral in Equation A.2 as:

$$\begin{aligned} \int_0^1 [d(\psi_i, \lambda_l)d(\psi_{i'}, \lambda_l)]^2 d\lambda_l &= \int_0^{1/2} [y(y - \Delta_{ii'})]^2 dy \\ &+ \int_{1/2}^{1/2+\Delta_{ii'}} [(1-y)(y - \Delta_{ii'})]^2 dy \\ &+ \int_{1/2+\Delta_{ii'}}^1 [(1-y)(1-y + \Delta_{ii'})]^2 dy \end{aligned} \quad (\text{A.3})$$

The first integral in Equation A.3 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 A.3 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 A.3 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:

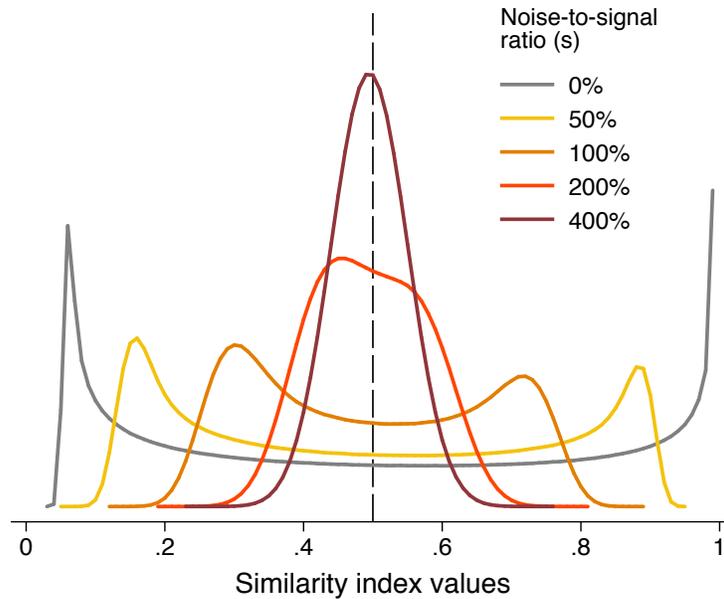
$$\int_0^1 [d(\psi_i, \lambda_l)d(\psi_{i'}, \lambda_l)]^2 d\lambda_l = \frac{-80\Delta_{ii'}^4 + 160\Delta_{ii'}^3 - 80\Delta_{ii'}^2 + 6}{480} = \frac{1}{180} - \frac{1}{6} \left( \Delta_{ii'} - \Delta_{ii'}^2 \right)^2$$

Plugging back calculated numerator and denominator into Equation A.1, we obtain:

$$\begin{aligned} \text{corr}\{r_i, r_{i'}\} &= \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}]}{E[(r_{il} - \bar{r}_i)^2|\psi_i]} = \frac{1/180 - (\Delta_{ii'} - \Delta_{ii'}^2)^2 / 6}{1/180} \\ &= 1 - 30 \left( \Delta_{ii'} - \Delta_{ii'}^2 \right)^2 = 1 - 30 \left( d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'}) \right)^2 \end{aligned}$$

Then the expected similarity between industries  $i$  and  $i'$  is:

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 = 1 - 15 \left( d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'}) \right)^2 \quad (\text{A.4})$$



**Figure A.1:** Similarity distributions for different noise levels.

### A.2.2 Simulating the similarity distribution with noise.

We tested the validity of expression in Equation A.4 for different noise-to-signal levels. As shown in Figure A.1, when the noise level is 0, we have two peaks, one at  $1/16$  (minimum value) and 1 (maximum value). As the amount of noise increases, the peaks become closer and closer, and finally merge to form a single peak around 0.5. This is expected because the coefficients of correlating pure uniform noise would eventually take on a normal distribution as the noise term (which is normally distributed) becomes larger compared to the signal term.

### A.2.3 General case with an n-dimensional sphere and an arbitrary functional form.

Instead of our specific choice of circle space and the functional form, we will present a more general case below. Again, we will assume that the efficiency with which industry  $i$  functions in location  $l$  depends on the distance between the requirements of industry  $i$  and endowments of location  $l$ . Suppose the requirements of the industry  $i$  are characterized by a parameter  $\psi_i$ , which is a point on the unit sphere in  $\mathbb{R}^n$  denoted by  $\mathbb{S}$ . The choice of unit sphere in arbitrary dimensions will make the calculations easy but it is not critical. Because of the symmetry of the surface of the sphere, there is no special points on the sphere if every attribute is uniformly distributed. Similarly, the endowments of location  $l$  is characterized by a parameter  $\lambda_l$ , also on  $\mathbb{S}$ . The output intensity of industry  $i$  in location  $l$  will depend on the congruity between the requirements of the industry,  $\psi_i$ , and the endowments of the location,  $\lambda_l$ . More concretely,

$$r_{il} = f(d(\psi_i, \lambda_l)) \quad (\text{A.5})$$

where  $d$  is the distance on  $\mathbb{S}$ , and  $f$  is a differentiable, decreasing function of the distance, such that  $f(0) = 1$  and  $f(1) = 0$  (After normalizing the maximum distance on the sphere to 1, hence the name unit sphere). As can be observed, output intensity will be maximized when  $\psi_i = \lambda_l$ ; in the opposite case, where  $\psi_i$  and  $\lambda_l$  are furthest away from each other, output would be close to zero. I

Again, we define as the product space similarity matrix  $\phi_{ii'}$  between two industries  $i$  and  $i'$  as the scaled Pearson correlation between  $r_i$  and  $r_{i'}$  across all locations:

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 \quad (\text{A.6})$$

where corr is defined as

$$\text{corr}\{r_i, r_{i'}\} = \frac{\sum_l (r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})}{\sqrt{\sum_l (r_{il} - \bar{r}_i)^2 \sum_l (r_{i'l} - \bar{r}_{i'})^2}}$$

Since each  $\psi_i$  and  $\lambda_l$  are independently and uniformly distributed, using law of large numbers, the sums in the correlation expressions can be converted to expectation values, namely:

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}]}{\sqrt{E[(r_{il} - \bar{r}_i)^2|\psi_i]E[(r_{i'l} - \bar{r}_{i'})^2|\psi_{i'}]}}$$

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

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}]}{E[(r_{il} - \bar{r}_i)^2|\psi_i]} \quad (\text{A.7})$$

Let's first start with the denominator. The denominator can be written as:

$$E[(r_{il} - \bar{r}_i)^2|\psi_i] = E[r_{il}^2|\psi_i] - (E[r_{il}|\psi_i])^2$$

Using the functional form in A.5, we can calculate:

$$\begin{aligned} E[r_{il}^2|\psi_i] &= \int_{\lambda_l \in \mathcal{S}} (f(d(\psi_i, \lambda_l)))^2 d\lambda_l \\ E[r_{il}|\psi_i] &= \int_{\lambda_l \in \mathcal{S}} f(d(\psi_i, \lambda_l)) d\lambda_l \end{aligned}$$

Similarly, we can write the numerator in A.7 as:

$$E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}] = E[r_{il}r_{i'l}|\psi_i, \psi_{i'}] - (E[r_{il}|\psi_i])^2$$

with:

$$E[r_{il}r_{i'l}|\psi_i, \psi_{i'}] = \int_{\lambda_l \in \mathcal{S}} f(d(\psi_i, \lambda_l))f(d(\psi_{i'}, \lambda_l))d\lambda_l$$

Using Equation A.7, we can write:

$$1 - \text{corr}\{r_i, r_{i'}\} = \frac{E[r_{il}^2|\psi_i] - E[r_{il}r_{i'l}|\psi_i, \psi_{i'}]}{E[r_{il}^2|\psi_i] - (E[r_{il}|\psi_i])^2}$$

Denominator of this expression does not depend on particular choices of  $\psi_i$  and  $\psi_{i'}$ . For

the sake of brevity, we will call it  $\sigma^2$ . We can write the numerator as:

$$\begin{aligned} E[r_{ii}^2|\psi_i] - E[r_{ii}r_{i'i}|\psi_i, \psi_{i'}] &= \int_{\lambda_i \in \mathbb{S}} (f(d(\psi_i, \lambda_i)))^2 d\lambda_i - \iint_{\lambda_i \in \mathbb{S}} f(d(\psi_i, \lambda_i))f(d(\psi_{i'}, \lambda_i))d\lambda_i \\ &= \iint_{\lambda_i \in \mathbb{S}} f(d(\psi_i, \lambda_i)) [f(d(\psi_i, \lambda_i)) - f(d(\psi_{i'}, \lambda_i))] d\lambda_i \end{aligned}$$

Let's move the origin to  $\psi_i$  and define  $\psi_{i'} = \psi_i + \Delta_{ii'}$ . Hence, the expression above can be written as a function of  $\Delta_{ii'}$  as:<sup>24</sup>

$$g(\Delta_{ii'}) = \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) [f(d(0, \lambda_i)) - f(d(\Delta_{ii'}, \lambda_i))] d\lambda_i$$

For a small increase,  $\delta_{ii'}$ , then we obtain:

$$g(\Delta_{ii'} + \delta_{ii'}) = \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) [f(d(0, \lambda_i)) - f(d(\Delta_{ii'} + \delta_{ii'}, \lambda_i))] d\lambda_i$$

Hence:

$$g(\Delta_{ii'} + \delta_{ii'}) - g(\Delta_{ii'}) = \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) [f(d(\Delta_{ii'} + \delta_{ii'}, \lambda_i)) - f(d(\Delta_{ii'}, \lambda_i))] d\lambda_i$$

If we divide both sides with  $\delta_{ii'}$ , we get:

$$\frac{g(\Delta_{ii'} + \delta_{ii'}) - g(\Delta_{ii'})}{\delta_{ii'}} = \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) \frac{f(d(\Delta_{ii'} + \delta_{ii'}, \lambda_i)) - f(d(\Delta_{ii'}, \lambda_i))}{\delta_{ii'}} d\lambda_i$$

The left hand side is the definition of derivative. Because of the symmetry in the distance function, we can rewrite the fraction under the integral as:

$$\begin{aligned} \frac{dg(\Delta_{ii'})}{d\Delta_{ii'}} &= \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) \frac{f(d(\Delta_{ii'}, \lambda_i - \delta_{ii'})) - f(d(\Delta_{ii'}, \lambda_i))}{\delta_{ii'}} d\lambda_i \\ &= - \iint_{\lambda_i \in \mathbb{S}} f(d(0, \lambda_i)) \frac{df(d(\Delta_{ii'}, \lambda_i))}{d\lambda_i} d\lambda_i \end{aligned}$$

<sup>24</sup>Note that, for the one-dimensional circle with circumference 1, this function is:

$$g(\Delta_{ii'}) = \frac{1}{6} \left( \left( \Delta_{ii'} - \Delta_{ii'}^2 \right)^2 \right)$$

$d(\Delta_{ii'}, \lambda_l)$  increase as  $\lambda_l$  increase in the direction of  $\Delta_{ii'}$ . Since  $f$  is a strictly-decreasing function, the derivative within the integral is negative. Since all other terms are positive, the integral becomes negative. And because of the negative sign in front of the integral, we obtain:

$$\frac{dg(\Delta_{ii'})}{d\Delta_{ii'}} > 0.$$

Following Equation A.5, the similarity measure is:

$$\phi_{ii'} = 1 - \frac{g(\Delta_{ii'})}{2\sigma^2}.$$

Therefore, we prove that the  $\phi_{ii'}$  strictly decreases as the distance between  $\psi_i$  and  $\psi_{i'}$  increases. This result implies that our similarity measure captures the distance, since this measure is a strictly decreasing function of the distance.

## A.3 Motivation Based on Factor Content of Production

### A.3.1 Relation to the Heckscher-Ohlin Model

This paper is related to the controversy surrounding the Leontief Paradox which has been seen as a major handicap of the Heckscher-Ohlin trade models. 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.<sup>25</sup> In most cases, it was not possible to list all factors related to the production 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.

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.

---

<sup>25</sup>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 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.

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:

$$AY = F \tag{A.8}$$

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  $r_{i,l}$  represents location  $l$ 's output in industry  $i$ .
- $F = N_f \times N_l$  is a matrix where  $F_{f,l}$  represents location  $l$ 's endowments of factor  $f$ .

From an empirical point of view, we can only observe  $Y$  is 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 A.8 by the  $B$  matrix, we obtain:

$$Y = BF \tag{A.9}$$

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.

### A.3.2 Similarities between the factor requirements of two industries

We will now derive a measure of input similarity of two industries, using Equation A.9. 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:

$$r_{il} = \sum_f B_{if} F_{fl} \quad (\text{A.10})$$

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

$$r_i = B_i F \quad (\text{A.11})$$

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 A.11, average production of industry  $i$  can be calculated as:

$$\bar{r}_i = \frac{\sum_l r_{il}}{N_l} = \sum_f B_{if} \frac{\sum_l F_{fl}}{N_l} = \sum_f B_{if} \bar{F}_f \quad (\text{A.12})$$

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

$$r_i - \bar{r}_i = B_i (F - \bar{F}) \quad (\text{A.13})$$

where  $\bar{F}$  is a  $N_f \times N_l$  matrix that repeats in each row  $f$  the average endowment of the world in that factor  $\bar{F}_f$ . Using Equation A.13, we can relate the observed covariance of the rows of the  $Y$  matrix to those of the unobserved  $B$  matrix:

$$(r_i - \bar{r}_i)(r_{i'} - \bar{r}_{i'})^t = B_i (F - \bar{F})(F - \bar{F})^t B_{i'}^t \quad (\text{A.14})$$

$C \equiv (F - \bar{F})(F - \bar{F})^t$  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:

$$C = U \Sigma U^t \quad (\text{A.15})$$

where  $U$  is a unitary matrix formed by the eigenvectors of  $C$  and  $\Sigma$  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 A.16 as

$$(r_i - \bar{r}_i)(r_{i'} - \bar{r}_{i'})^t = \sum_f \tilde{B}_{if} \tilde{B}_{i'f}^t \sigma_f \quad (\text{A.16})$$

where  $\sigma_f$  is the  $f^{\text{th}}$  (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

$$\sum_f \tilde{B}_{if} \tilde{B}_{i'f}^t \sigma_f = \sigma \tilde{B}_i \tilde{B}_{i'}^t = \sigma B_i U U^t B_{i'}^t = \sigma B_i B_{i'}^t \quad (\text{A.17})$$

Dividing both sides of Equation A.17 by the standard deviation of  $r_i$  and  $r_{i'}$ , we can relate the correlation of the rows of the  $Y$  matrix to elements of the  $B$  matrix:

$$\text{corr}\{r_i, r_{i'}\} = \frac{(r_i - \bar{r}_i)(r_{i'} - \bar{r}_{i'})^t}{\sigma_{r_i} \sigma_{r_{i'}}} \approx \frac{\sigma}{\sigma_{r_i} \sigma_{r_{i'}}} B_i B_{i'}^t \quad (\text{A.18})$$

where  $\text{corr}$  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  $i'$ :

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 \quad (\text{A.19})$$

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

### A.3.3 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  $r_l$  and  $F_l$  as the  $l^{\text{th}}$  column vectors of  $Y$  and  $F$  matrices respectively, the output of a location is related to its factor endowments by:

$$r_l = B F_l \quad (\text{A.20})$$

Note that our calculations in Section 2.1.1 can be replicated here because if we take the transposes of both sides in Equation A.20, we will arrive to an expression similar to Equation A.11. Assuming that the columns of  $B$  matrix are independently and identically distributed, we can write (akin to Equation A.18):

$$\text{corr}\{r_l, r_{l'}\} = \frac{(r_l - \bar{r}_l)^t (r_{l'} - \bar{r}_{l'})}{\sigma_{r_l} \sigma_{r_{l'}}} \approx \frac{\sigma'}{\sigma_{r_l} \sigma_{r_{l'}}} F_l^t F_{l'} \quad (\text{A.21})$$

where  $\bar{r}_l$  is the average production of location  $l$ ,  $\sigma_{r_l}$  is the standard deviation of  $r_l$ ,  $\sigma'$  is the diagonal of the covariance matrix ( $(B - \bar{B})^t (B - \bar{B}) \approx \sigma' I_{N_f}$ ). 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:

$$\phi_{ll'} = (1 + \text{corr}\{r_l, r_{l'}\})/2 \quad (\text{A.22})$$

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

### A.3.4 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  $\hat{r}$ ,  $\hat{A}$  and  $\hat{F}$  matrices such that  $\hat{r}_{il} = r_{il}/(s_i s_l)$ ,  $\hat{A}_{fi} = s_i A_{fi}$  and  $\hat{F}_{fl} = F_{fl}/s_l$  then

$$\hat{A} \hat{r} = \hat{F} \quad (\text{A.23})$$

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:

$$r_{il} = X_{il} + C_{il} \quad (\text{A.24})$$

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:

$$r_i - \bar{r}_i = (X_i - \bar{X}_i) + (C_i - \bar{C}_i) \quad (\text{A.25})$$

Assuming homothetic preferences worldwide, and normalizing each industry element by its size, we can assume that  $C_i = \bar{C}_i$ . 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  $X$ :

$$r_l - \bar{r}_l = (X_l - \bar{X}_l) + (C_l - \bar{C}_l) \quad (\text{A.26})$$

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

### A.3.5 Simulating the estimators on an HOV toy model

We test the effectiveness of our estimators of  $r_{il}$  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  $r_{il}$  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 A.18. We randomly select 5,000  $A^{-1}$

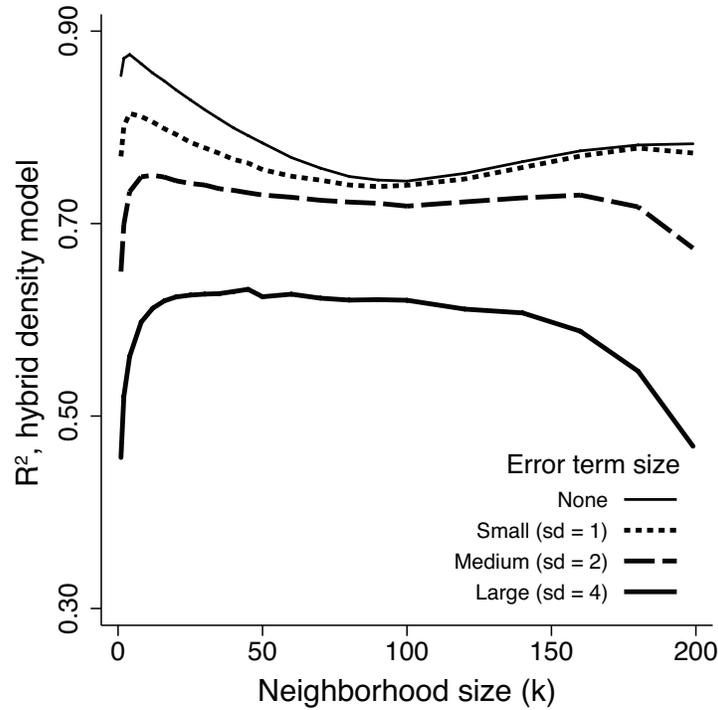
and  $F$  matrices and test the validity of this equation. We note that the random selection of both matrices simultaneously puts no inherent structure into these matrices and in reality we expect to observe more structures matrices. Even in the random case, the correlation between the actual and estimated numbers exhibit is  $0.532 \mp 0.014$ . 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 A.21 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 A.2, the uppermost line shows the effect of neighborhood size on the  $R^2$ . We see that the highest  $R^2$  value is found at  $k = 4$ .

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,  $\tilde{r}_{il}$ , is affected by a random error term,  $\varepsilon_{il}$ , with a normal distribution around a mean of zero:

$$\tilde{r}_{il} = r_{il} + \varepsilon_{il} \tag{A.27}$$

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  $r_{il}$  by averaging the observed, noisy output  $\tilde{r}_{il}$  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  $r_{il}$  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



**Figure A.2:** Simulation of association between underlying output and hybrid density model, by size of neighborhood and noise level.

the standard deviation of  $r_{il}$ , respectively.

In Figure A.2, 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.