Implied Comparative Advantage

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

CID Faculty Working Paper No. 276

January 2014
Revised July 2020

Working Papers

Center for International Development
at Harvard University

Implied Comparative Advantage [*]

Ricardo Hausmann **César A. Hidalgo** **Daniel P. Stock**
Muhammed A. Yıldırım[†]

May 2020

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 evolution of comparative advantage. We build a simple Ricardian-inspired model and show that hidden information on inter-industry and inter-location relatedness can be captured by simple correlations between the observed structure of industries across locations, or the structure of locations across industries. Using this information from related industries or related locations, we calculate a measure of *implied comparative advantage* and show that it 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). In each of these cases, 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

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

[†] Hausmann: Center for International Development at Harvard University (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), danielstock@cmu.edu (Stock), mayildirim@ku.edu.tr (Yıldırım).

1 Introduction

David Ricardo (1817)'s seminal theory predicts that locations benefit when they allocate their resources in the goods in which they have a comparative advantage, *i.e.*, those produced with a higher relative productivity. Yet these comparative advantage patterns are not random, nor are they set in stone; theories detailing the evolution of locations' productivity levels date back to the work of Marshall (1890) more than a century ago. Since then, many studies have highlighted the role of relatedness between sectors and relatedness between regions in the evolution of comparative advantage. Here, we take a complementary stance, giving evidence that these patterns of relatedness also reveal deeper information about the requirements of industries and endowments of locations. We then show how this information could be used to develop a measure of counterfactual or *implied* comparative advantage, and how such a measure helps explain changes in comparative advantage of locations over time.

According to the Ricardian theory of trade, the intensity of production of a location in an industry 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 and by its productivity in the industry relative to other locations. 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). Yet these models can only infer the relative productivity of a location in a product if the location already makes the product.[1] This is an important void, as the emergence of new, modern industries is an essential component of economic development (Hausmann et al. 2007). In addition, current Ricardian models assume that the relative productivity parameters are uncorrelated across industries. This implies that the likely productivity of a country in motorcycle production, for example, is equally independent of whether it currently has comparative advantage in car-making or in coffee. We provide evidence that appears to contradict this.

In this paper we extend the neo-Ricardian models to address these issues. In our simplified model, we assume that the comparative advantage is determined by the distance between factor endowments of locations and factor requirements of industries. We then illustrate how, given any two industries, the distance between their factor requirements

[1] 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."*

can be linked to the correlation between their output levels (in terms of their respective patterns of comparative advantage across locations). That is, two industries with very similar factor requirements will tend to have similar levels of comparative advantage (in each location). Likewise, smaller differences between the factor endowments of locations are translated into higher levels of correlations between their respective comparative advantage patterns as well. If our model is correct, then deep information on industries and locations can be intuited from surface-level patterns in comparative advantage. In particular, it would imply that the comparative advantage of an industry in a location (or ”industry-location”) can be estimated from the comparative advantage of highly correlated industries, or highly correlated locations. This is true even for industry-locations that are currently absent or unobserved.

We then propose how to construct such estimates. Unlike other predictive approaches in the diversification and complexity literature, we build a proxy that expresses the expectations of an underlying factors model, that is, the implied comparative advantage of an industry-location. We then extend our theoretical model to show how regression residuals from such proxies would be expected to predict future changes in comparative advantage, among industry-locations that already exist or those that have yet to emerge.

Finally, we use a variety of datasets to construct these proxies and verify their predictive power. First, we show that our measures are highly significant predictors of international export flows – both present-day export patterns and industry-location export growth. Next, we apply our model at the subnational level, using data from the US, India and Chile.[2] With this data, we obtain similar results when constructing our implied comparative advantage measures using the wage bill, employment or the number of establishments of industry-locations. Our results also 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. Extending the trade models to make predictions on the extensive margin could be crucial for shaping policy discussions, given the special importance of the emergence of new industries.

Together, these results appear to confirm the predictions of our model: that (1) information on hidden endowments and requirements can be recovered from an analysis of the realized economic structures (*i.e.*, *observed* comparative advantage), (2) this information can be used to construct a proxy of *implied* comparative advantage, and (3) the

[2] We are not the first to apply international trade models to a subnational setting; see, for example, Davis and Dingel (2014), Costinot et al. (2016), and Caliendo et al. (2017). 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.

present-day gap between implied comparative advantage and observed comparative advantage is associated with long-term changes in observed comparative advantage.

The rest of the paper is structured as follows. Section 2 gives an overview of the related literature. Section 3 provides the basic model behind our findings. Section 4 discusses the data and methodology used to build our variables. Section 5 presents our main empirical results: explaining the current structure of industry-locations, and exploring links with future growth. Section 6 tests some direct implications of our model, and evaluates the alternative explanations and robustness of our results. In Section 7, we discuss the implications of our findings and conclude.

2 Related Literature

This paper relates to several strands of literature, given that it covers international trade, growth, and subnational settings (cities and regions). It most directly builds on Hausmann and Klinger (2006), Hidalgo et al. (2007), Hausmann and Klinger (2007) and Bahar et al. (2014), developing an underlying theoretical foundation to the empirical patterns described in those papers. It also expands past literature by exploring the intensive margin of industry-location growth (in addition to appearance and disappearance), and refining the measures they use.

Other research explores the theoretical underpinnings of diversification. Boschma and Capone (2015) analyze the interaction between relatedness and institutions and find 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 subnational 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 agglomerative forces as well. Delgado et al. (2010, 2015) and Porter (2003) use US subnational data to explain employment growth at the city-industry level, using the presence of related industry clusters. Lu et al. (2016) explore 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) survey 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 measures that we use capture more than the Marshallian forces.

In the international context, 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) assumes that the productivity parameters are drawn from a Frechét distribution, except for a common national productivity parameter. Costinot et al. (2012) relaxes 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 we document patterns of positive and negative correlation across export industries in the same country. Finally, our approach has the advantage of being able to estimate relative productivities for industries that currently have zero (or unobserved) output. Previous Ricardian literature, however, cannot infer relative productivities of industries that do not yet exist.[3]

Our approach uses two-dimensional industry-location matrices to explain the evolution of revealed comparative advantage. The economic complexity literature building 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 (2019). These measures can also be used to model new product appearances in the context of evolution of complexity. Nevertheless, they do not aim to model or predict industry-location-level production patterns as we do here.

Finally, the measures we derive are similar to the collaborative filtering recommenda-

[3] An exception is Costinot et al. (2016), who estimate implied or counter-factual productivity for agricultural industries using agronomic models and data. This requires detailed data and knowledge of agricultural production functions and, hence, cannot easily be extended to other settings.

tion models in computer science. 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 items based on ratings by similar users (Resnick et al. 1994). But these techniques never ask *why* a pair of consumers or products might have correlated preferences. Here, we derive a theoretical rationale for their logic.

3 Model

In this section, we use a modified Ricardian framework to show how patterns in the *observed* or *revealed* comparative advantage of locations can contain information on their "true" comparative advantage, *i.e.*, the hidden match between the requirements of industries and the ability of locations to meet those requirements.

To begin, we first need a definition of revealed comparative advantage. Let's denote the output of an industry i in a location l with y_{il}. It follows that the total output of an industry is $Y_i \equiv \sum_l y_{il}$. Now, let us construct a counterfactual industry-location output estimate, $\hat{y}_{il}$, without any differences in comparative advantage across locations. In this no-advantage world,[4] each location would produce its "fair share" in each industry; a fair share based on population, for example, would be:

$$\hat{y}_{il} = s_l Y_i \tag{3.1}$$

where s_l is location l's share of total population ($s_l = population_l / population_{world}$). One could also calculate a fair share using the location's proportion of global output, exports, value added or employment.

Since $\hat{y}_{il}$ is our representation of a world structure without differences in productivities, then we can define our comparative advantage term, r_{il}, as the ratio between that no-advantage world and the real world:

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

Taking logs and re-arranging terms gives a way to express all industry-location output:

$$\log(y_{il}) = \log(r_{il}) + \log(Y_i) + \log(s_l) \tag{3.2}$$

In this paper, we use s_l to be the population share of the location. In the international

[4] We assume there are no economies of scale and individuals have identical preferences in all locations.

context, y_{il} is the exports of country l in industry i.[5] Alternatively, if we use y_{il} to be the number of employees in industry i in location l and s_l is the share of employment of the location in the country, we arrive at the widely-used Location Quotient (LQ) measure.

In a sense, Equation 3.2 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 size dynamics (s_l), and a portion that is specific to the interaction of locations and industries (r_{il}). In our empirical analysis, we will be focusing in this interaction term.[6]

3.1 Modeling comparative advantage

Having defined our measure of industry-location comparative advantage, we can now model how these values 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 $\mathbb{S}$. Suppose the requirements of the industry i are characterized by a parameter $\psi_i \in \mathbb{S}$ and the endowments of location l is characterized by a parameter $\lambda_l \in \mathbb{S}$. The output intensity of industry i in location l (r_{il}) will depend on some function of the distance between ψ_i and λ_l:

$$r_{il} = f\left(d(\psi_i, \lambda_l)\right) \tag{3.3}$$

where d is the distance metric on our compact metric space $\mathbb{S}$, and f is a strictly decreasing function of the distance, such that $f(0) = 1$ and $f(d_{max}) = 0$, where d_{max} is the maximum distance in $\mathbb{S}$. In other words, r_{il} is increasingly large as the industry requirement and location endowment are closer together.[7]

In reality, we are not able to observe $\psi_i$and λ_l directly – they are hidden from the observer. However, we do observe the r_{il}, and can in fact use then to glean information

[5] If we take s_l to be the share of the country in world trade, then r_{il} becomes Balassa (1964)'s Revealed Comparative Advantage (RCA) measure. See the Appendix for our results using RCA.

[6] 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?"

[7] We introduce a more structural model in the Appendix, in which labor productivity is the consequence of the requirements and availability of multiple factors of production. In this setting, based on the canonical Heckscher-Ohlin-Vanek trade model, we reproduce the same key results as those given here.

about the underlying ψ_i and λ_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 ψ_i and $\psi_{i'}$. By the same token, the difference in the same industry across two locations, r_{il} and $r_{il'}$, would be an increasing function of the difference in the λ_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 r-values in the same locations) or one location to another (by comparing their r-values 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 industry similarity matrix $\phi_{ii'}$ between two industries i and i' as the scaled Pearson correlation between[8] r_i and $r_{i'}$ across all locations:

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 \tag{3.4}$$

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

$$\phi_{ll'} = (1 + \text{corr}\{r_l, r_{l'}\})/2 \tag{3.5}$$

We want to show that ϕ captures the distance between two industries' requirements or two locations' endowments in our compact space $\mathbb{S}$. That is, without knowing what ψ_i and $\psi_{i'}$ are, we would like to infer the distance between them using $\phi_{ii'}$. In fact, it can be shown (Section A.1.1 of Appendix) that if we model our space $\mathbb{S}$ as a unit-sphere in an arbitrary number of dimensions ($\mathbb{R}^n$), and use any function f satisfying the rules described above, it follows that $\phi_{ii'}$ ($\phi_{ll'}$) strictly decreases as distance between ψ_i and $\psi_{i'}$ (λ_l and $\lambda_{l'}$) increases:

$$\frac{\partial \phi_{ii'}}{\partial d(\psi_{i'}, \psi_i)} < 0, \qquad \frac{\partial \phi_{ll'}}{\partial d(\lambda_{l'}, \lambda_l)} < 0. \tag{3.6}$$

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 space, where ψ and λ are points on a

[8] As a notational convention going forward, when we skip an index subscript, it means that it becomes a vector over the skipped variable. In this case, r_i is the vector whose elements are r_{il}.

unit-sphere of one dimension, *i.e.*, the unit circle, $\mathbb{U}$.[9] On $\mathbb{U}$, output intensity is maximized when $\psi_i = \lambda_l$; in the opposite case, where ψ_i and λ_l are on antipodal sides of the circle (and distance is 0.5), output is zero. In addition, we now assign a specific functional form for f, in order to calculate actual r_{il} values:

$$f\left(d(\psi_i, \lambda_l)\right) = 1 - 4d^2(\psi_i, \lambda_l). \tag{3.7}$$

If we assume that ψ_i and λ_l are uniformly distributed on $\mathbb{U}$, then we can derive a closed form expression for the expected value of the $\phi_{ii'}$ ($\phi_{ll'}$) as a monotonic function of the distance between ψ_i and ψ_i' (λ_l and $\lambda_{l'}$) (see Appendix section A.1.2):

$$\phi_{ii'} = 1 - 15\left(d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'})\right)^2, \qquad \phi_{ll'} = 1 - 15\left(d(\lambda_l, \lambda_{l'}) - d^2(\lambda_l, \lambda_{l'})\right)^2 \tag{3.8}$$

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.[10] If ψ and λ are distributed uniformly, we can expect to find a wide range of ϕ values.

Thus, we can conclude that our similarity measures – built from observable industry-location information – are directly linked to the difference between two industries' unobserved factor requirements, or two locations' unobserved factor endowments.

3.2 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 in a second location, l'; this means that $\phi_{ll'} \approx 1$. Based on equations 3.8 and 3.7, 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 3.8 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

[9] 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.

[10] Using this form for f, the minimum similarity value 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 to get perfect anti-correlation on the unit circle. Even in that case, the points at 90° and 270° will have similar values of comparative advantage.

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 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.[11] Thus, if we base our proxy on the single most related industry or location, we may be introducing a large error in our prediction. In fact, using a larger number of comparators would improve the odds that we are including locations whose endowments deviate from λ_l in opposite directions, *i.e.*, places whose differences from l will "cancel out" each other.[12]

Second, it is unlikely that one 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. 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 ψ_i and the technological ability of the location λ_l. Let us assume that the comparative advantage of each industry-location has deviated from this underlying value because of a disturbance term ε_{il}:

$$\tilde{r}_{il} = r_{il} + \varepsilon_{il}, \tag{3.9}$$

What could cause this disturbance? 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, we no

[11] 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 = 0.690). Industries, however, are likelier to have a highly similar best comparator (median = 0.902).

[12] Consider a location l with two highly similar comparators, l' and l''. In cases where $\lambda_{l'} > \lambda_l > \lambda_{l''}$ or $\lambda_{l''} > \lambda_l > \lambda_{l'}$, the average of $r_{il'}$ and $r_{il''}$ will be closer to r_{il} than either $r_{il'}$ or $r_{il''}$ individually. For example, if we set $\{\lambda_l, \lambda_{l'}, \lambda_{l''}\}$ to $\{0.75, 0.77, 0.72\}$ and ψ_i equal to 0.44, then we get $\{r_{il}, r_{il'}, r_{il''}\} = \{0.62, 0.56, 0.69\}$; the mean of $r_{il'}$ and $r_{il''}$ – 0.63 – is much closer to r_{il} than either $r_{il'}$ or $r_{il''}$ are.

longer see the *true* comparative advantage r_{il}; we can only witness *observed* comparative advantage $\tilde{r}_{il}$. Again, our comparators will come to rescue: by integrating information from several meaningful comparators, we might be able to predict r_{il}. Since we do not observe the sign of the disturbance term, including a larger number of comparators ensures there are enough positively biased values to offset negatively biased ones.

Going forward, for any location l, we can formally define our expected value of its true r_{il} value using the following approach. First, we use the similarity parameters $\phi_{ll'}$ to build the set of its k most similar comparator locations. Next, we take a weighted mean of the $\tilde{r}_{il'}$ values of each of the locations in this set, with the $\phi_{ll'}$ serving as the weights. We refer to this proxy for implied comparative advantage as *country space density*:

$$\widehat{r}_{il}^{[L]} = \sum_{l' \in L_l(k)} \frac{\phi_{ll'}}{\sum\limits_{l'' \in L_l(k)} \phi_{ll''}} \tilde{r}_{il'} \tag{3.10}$$

with the set $L_l(k)$ defined as:

$$L_l(k) = \{l' | Rank\left(\phi_{ll'}\right) \leq k\} \tag{3.11}$$

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

$$\widehat{r}_{il}^{[I]} = \sum_{i' \in I_i(k)} \frac{\phi_{ii'}}{\sum\limits_{i'' \in I_i(k)} \phi_{ii''}} \tilde{r}_{i'l} \tag{3.12}$$

where set $I_i(k)$ contains the k nearest neighbors of industry i:

$$I_i(k) = \{i' | Rank\left(\phi_{ii'}\right) \leq k\} \tag{3.13}$$

We refer to this variable as *product space density*.[13] As the two density measures are symmetric, we can expect that they capture some non-overlapping information; that is, the best proxy may be a hybrid of both of them.

We do not particularly impose a structure on the disturbance term in Equation 3.9, ε_{il}. The density measures rely on taking an average to smooth out the effects of these terms. Hence, for that to be true, the mean of ε_{il} terms should be 0 and the variance of these

[13] Our density variable is in fact similar to the product space density first proposed by Hausmann and Klinger (2006), adapted to describe continuous inputs.

terms should be well-defined and finite, as in the central limit theorem. If the variance is not bounded, then there might be some extreme values skewing the results. If the mean is not 0, then the density measures would give results biased towards this mean value; this would be largely captured by the regression coefficient and/or constant term.[14]

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 ($k = N$). We later relax this constraint, testing the relationship between k and accuracy directly (Appendix A.3).

Note that neither $\widehat{r}_{il}^{[I]}$ nor $\widehat{r}_{il}^{[L]}$ contain direct information from $\tilde{r}_{il}$, meaning that we are not using an observation to proxy itself.[15] This also means that we can find non-zero 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.

3.3 Predicting trends towards true comparative advantage

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

In Equation 3.9 above, we introduce a disturbance term, ε_{il}, that captures locations' tendency to deviate from their *true* comparative advantage, allocating resources suboptimally. As such, let us assume that these deviations are unsustainable, and will diminish over long periods of time, *i.e.*, $\lim_{t\to\infty} \varepsilon_{il,t} = 0$.[16] 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:

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

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

[14] Likewise, if the mean of disturbance term varies by industry or location, this effect would be largely absorbed by industry or location fixed effects (if not the controls). The only truly troubling scenario would be persistent industry-location bias in the disturbance, e.g. machinery exports are artificially high in European comparator countries but not in Asian comparator countries.

[15] Strictly speaking, $\tilde{r}_{il}$ is still included in the similarity index calculation; section A.2.2 uses a cross-validation approach to verify that completely excluding $\tilde{r}_{il}$ from density does not impact our results.

[16] To declutter our notations, we only use time indices (t) when we explicitly study the time dimension.

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 ($\widehat{r}_{il,t_0} - \tilde{r}_{il,t_0}$) and *changes over time* in observed comparative advantage ($\tilde{r}_{il,t_1} - \tilde{r}_{il,t_0}$):

$$\tilde{r}_{il,t_1} - \tilde{r}_{il,t_0} = \beta(\widehat{r}_{il} - \tilde{r}_{il,t_0}), \qquad \beta > 0 \tag{3.14}$$

3.4 Hypotheses

We can now summarize the theoretical hypotheses we wish to test.

H1: If the output of an industry-location can be estimated based on highly similar industries and locations, then our measures of *implied* comparative advantage should be strongly and positively associated with measures of *observed* comparative advantage.

H2: If industry-locations are displaced from their *true* Ricardian comparative advantage, then (**H2a**) information regarding the true (non-displaced) comparative advantage will still be *implied* by highly similar industries and locations, and (**H2b**) our measures of this *implied* comparative advantage will give a better estimate of true comparative advantage than predictions using observable (displaced) output levels.

H3: If the displacement is diminishing over time, with each industry-location's growth tending towards its true comparative advantage, then (**H3a**) the difference between present output and future output should be correlated with the difference between present output and our implied comparative advantage measures. Furthermore, (**H3b**) the explanatory power of this relationship should increase for longer-term growth, as the real world moves closer to the Ricardian ideal over time.

H4: Since information on true comparative advantage can be estimated even for industries currently absent from a location, then (**H4a**) the implied comparative advantage measures will be predictive of which new industries will emerge (or fail to emerge). Conversely, (**H4b**) industries that are not supported by the current economic structure, *i.e.*, those that have low implied comparative advantage values, will disappear over time.

To investigate these hypotheses, we begin with a simulation of our theoretical model, verifying that the model performs as expected when we supply the underlying parameters.[17] We then test our remaining expectations empirically, using a variety of datasets.

[17] In particular, Hypothesis 2 is only verifiable in a simulated setting, since we never observe true comparative advantage in real life.

3.5 Simulating the theoretical 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 ψ_i and the λ_l along the unit circle $\mathbb{U}$. We then use these parameters to calculate the r_{il} values, using the functional form in Equation 3.7. Next, we model ε_{il} in in Equation 3.9 as normally distributed, with zero mean and variance equal to the variance of r_{il} (denoted by σ^2)[18] times a parameter s, the *noise-to-signal ratio* ($\varepsilon_{il} \sim N(0, s\sigma^2)$). We vary this noise-to-signal ratio, going from 0% (no error) to 100% (equal parts signal and noise) to 400% (four times more noise than signal). From the noisy output levels ($\tilde{r}_{il}$), we build the product space and country space densities, setting $k = \sqrt{N_l} = \sqrt{N_i} = 10$. We then measure the explanatory power of PS and CS density (and the mean of the two) by running linear regressions of the (noiseless) *true* comparative advantage values.

Table 1: Simulated explanatory power of observed and implied comparative advantage (mean R^2 from 5,000 simulations)

	Regression of *true* comparative advantage on…			
Noise-to-signal ratio	*observed* comparative advantage	*implied* comparative advantage		
		PS Density	CS Density	Hybrid
0%	1.000	0.982	0.982	0.990
25%	0.885	0.968	0.968	0.980
50%	0.640	0.927	0.927	0.955
100%	0.250	0.774	0.774	0.852
200%	0.040	0.370	0.370	0.490
400%	0.003	0.015	0.015	0.028

We carry out this exercise multiple times (5,000 simulations) and report the results in Table 1. First, we can verify the validity of Equation 3.8: on average, the absolute difference between the observed similarity value and the similarity value implied by the equations is less than 5%.[19] We can also check the distribution of the simulated similarity index values. As expected, there is a wide range of comparator pairs. This includes both

[18] We find that the average standard deviation of r_{il} converges to 0.298, so we use that value for σ^2.

[19] 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 locations and industries. In fact, increasing this number – from $N = 100$ to $N = 1000$ – decreases the error to 1.4%.

positively- and negatively-correlated comparators, distributed symmetrically about the midpoint (*i.e.*, zero correlation), though the exact shape of the distribution depends on the amount of noise added (see Figure A.1 in Appendix).

At this point, we can use the simulation to test the relation between our measures of *implied*, *observed* and *true* comparative advantage. Table 1 gives the mean R^2 values across our repeated simulations.[20] 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 (0.37 to 0.98).[21] 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, and combined together they estimate over 85% of the variance of the true values. 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, this implies that even if the underlying Ricardian dynamics have a relatively low explanatory power today, our density measures may be able to separate their message from the noise.[22]

4 Data and Methods

4.1 Data

We utilize international trade data to study the industry-location relationship at an international scale. Here we use UN COMTRADE data, downloaded from the Atlas of

[20] The R^2 values are highly consistent across the simulations, with standard errors of the means all less than a tenth of a percentage point.

[21] 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 the number of locations and industries is the same). Note also that the combination of the two densities is a stronger predictor than either one individually; this is an expected consequence of the law of large numbers, as we mentioned above.

[22] 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.

Economic Complexity.[23] Exports are disaggregated into product categories according the Harmonized System four-digit classification (HS4), for the years 1995-2016.[24] 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).[25] We then drop the 73 products with under $100 thousand in total world trade in 2005, and the miscellaneous code HS 9999. These restrictions reduce the sample to 119 countries and 1166 products, which in 2016 account for 92% of world trade and 93% of world population.

In addition to the international trade data, we use subnational data from three countries, namely the US, India and Chile. For the US, we use the County Business Patterns (CBP) database from 2003-2011.[26] 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 industries.[27] Our Chilean dataset comes from the Chilean tax authority, and includes the number of establishments based on tax residency for 334 municipalities and 681 industries, from 2005 to 2008 (see Bustos et al. (2012) for details). 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.[28] For all the datasets above, we include only industries and regions that have non-zero totals for all years. This approach effectively removes discontinued categories.

4.2 Constructing the model variables

For each dataset, we build the similarity and density measures as described above. 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

[23] The data and cleaning procedure can be found at: `https://atlas.cid.harvard.edu/about-data`

[24] 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.

[25] We also 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).

[26] During these years, two versions of NAICS (2002 and 2007) were used. If we extend the data to earlier or later years, we would need to convert an additional revision, which might introduce errors.

[27] The discrepancy between employment and establishment versus payroll sample sizes comes from the data suppression methods of Census Bureau. To protect the privacy of smaller establishments, the CBP occasionally discloses only the range of employment of an industry in a location, *e.g.*, 1 to 20 employees. In these censored cases, we use the 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.

[28] This is an earlier version of the data used in Asher and Novosad (2017).

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. For instance, for a country like Saudi Arabia, the price change in oil will result in changes in RCA in other industries even though the production levels in other industries do not change. We formally define the Revealed per-Capita Advantage (RpCA)[29] of location l in industry i as:

$$R_{il} = \frac{y_{il} / pop_l}{\sum_l y_{il} / \sum_l pop_l} \tag{4.1}$$

where y_{il} is the export, employment or payroll value, and pop_l is the population in location l. 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 3.4 and 3.5 below).[30] We do not normalize the data for the number of establishments.

At this point, we can use the normalized industry intensity values, R_{il}, to build the similarity indices defined above in Equations 3.4 and 3.5.[31] Tables 2 and 3 show the top ten most similar pairs of countries and products in the most recent year. We note that the most similar countries are those in the same geographic region, a phenomenon that can be explained by geological and climate effects as well as regional knowledge spillovers (Bahar et al. 2014). The list of most similar product pairs contains a mix of different categories, though the lower half of the list is dominated by machinery and electrical products. This matches the observation in Hausmann et al. (2014) that such industries are highly interconnected. That said, if we restrict the list to products in different Harmonized System chapters (panel b), the resulting pairs still seem highly intuitive.

Figure 1 shows the full distribution of the similarity index values. The right panel depicts location similarity values. This distribution is roughly symmetric, with a peak at the center (median = 0.509, mode at 0.49-0.50); this matches the "noisy world" predictions of the simulation (Figure A.1). However, the same cannot be said for the industry similarity values (left panel). The distribution of these values is also roughly symmetric, but

[29] Our results are robust to the use of standard RCA instead of RpCA. See Appendix for details

[30] 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 $R_{i,China} = pop_{World} / pop_{China} \approx 5$.

[31] 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
SDN	Sudan	TCD	Chad	0.867
CIV	Cote d'Ivoire	CMR	Cameroon	0.826
BGD	Bangladesh	KHM	Cambodia	0.795
COD	Congo, Dem. Rep.	COG	Congo, Rep.	0.788
COD	Congo, Dem. Rep.	ZMB	Zambia	0.781
JPN	Japan	KOR	Korea, Rep.	0.780
CIV	Cote d'Ivoire	GHA	Ghana	0.779
LTU	Lithuania	LVA	Latvia	0.765
CZE	Czech Republic	DEU	Germany	0.747
FIN	Finland	SWE	Sweden	0.745

distributed around a peak greater than 0.5 (median = 0.724, mode at 0.74-0.75). Industry similarity might have a higher than expected proportion of positive correlations because the distribution of the underlying industry technology parameters (ψ_i) could exhibit non-random patterns (i.e. not uniform-random, as our model assumes).[32] Incorporating such a structure is a promising area for future research (though out of scope for this paper).

Moving forward, we are particularly interested in the best comparators for each industry or location. Figure 1 illustrates the distribution of each industry or location's most similar comparator, as well as its $\sqrt{N}$ most similar 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 34 most similar comparators (interquartile range (IQR) = 0.81 to 0.90). In the right panel, we see that the most similar comparator locations tend to have somewhat lower similarity values. This includes the top 11 comparators (IQR = 0.59 to 0.65). 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 3.10 and 3.12, replacing the r_{il} with R_{il}. As before, we set the neigh-

[32] On a deeper level, this finding touches on the diversification versus specialization debate in the growth and trade literature. If locations stay narrowly specialized as they industrialize – replacing old industries with new ones – then we would expect to observe more negative correlations between industries' production locations. Instead, there appears to exist a *nested* structure of industry cooccurrence (Bustos et al. 2012), where new, more sophisticated goods are co-exported with more primitive ones (but not necessarily vice-versa). Our result supports this view.

Table 3: Most similar industry pairs, international trade, 2016

(a) All industry pairs

Industry i		Industry i'		Similarity
8484	Gaskets and similar joints	8485	Ships or boats propellers and blades	0.982
6204	Women's suits, not knit	6206	Women's shirts, not knit	0.980
8479	Specialized machines, mechanical appliances	8514	Industrial or laboratory electric furnaces	0.977
3921	Plastic plates, sheets, film, foil and strip	7326	Specialized articles of iron or steel	0.976
9030	Instruments for measuring electricity	9031	Optical or specialized measuring instruments	0.976
8483	Transmission shafts	8515	Electric soldering machines	0.975
8481	Thermostatically-controlled valve appliances	8484	Gaskets and similar joints	0.975
8207	Drilling, pressing and milling tools	8208	Knives and cutting blades for machines	0.975
8543	Specialized electrical machines and apparatus	9031	Optical or specialized measuring instruments	0.974
8536	Plugs, sockets, relays, other protective apparatus	8538	Electrical switch or protection components	0.974

(b) Industry pairs from different HS chapters

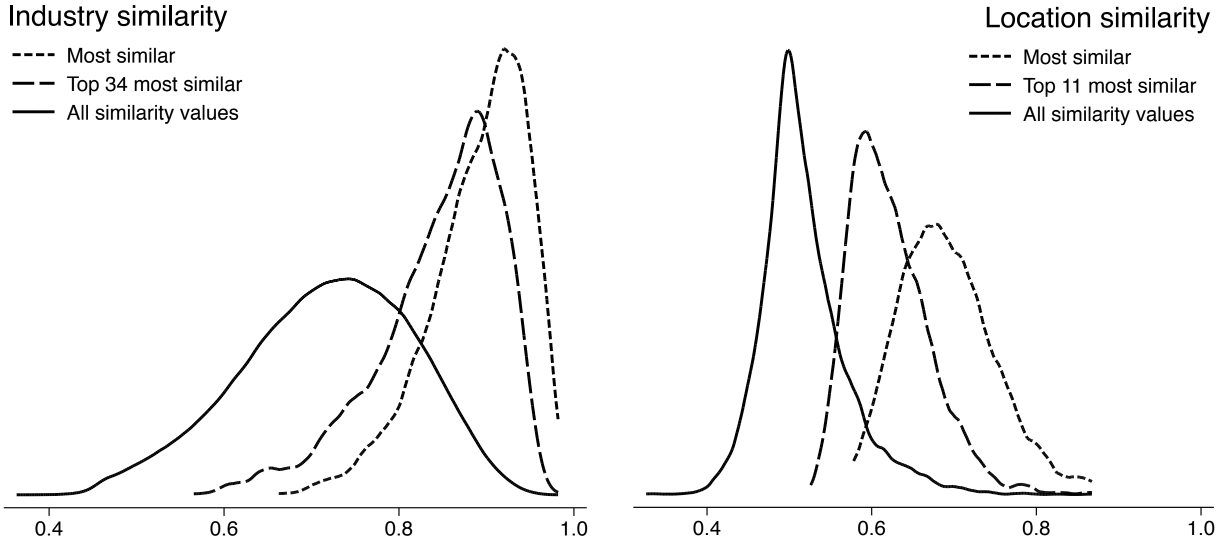
Industry i		Industry i'		Similarity
3921	Plastic plates, sheets, film, foil and strip	7326	Specialized iron or steel articles	0.976
8543	Specialized electrical machines and apparatus	9031	Optical or specialized measuring instruments	0.974
5911	Textile fabric for card clothing, technical use	3926	Specialized plastic articles	0.974
8208	Knives and cutting blades for machines	8466	Metalworking machine parts, accessories	0.974
7415	Copper nails, tacks and staples	8479	Specialized machines, mechanical appliances	0.972
7318	Screws, nuts, bolts, similar iron or steel articles	8466	Metalworking machine parts, accessories	0.970
8531	Electric sound or visual signaling apparatus	9031	Optical or specialized measuring instruments	0.970
8479	Specialized machines, mechanical appliances	9031	Optical or specialized measuring instruments	0.969
3921	Plastic plates, sheets, film, foil and strip	8419	Industrial heating and cooling machinery	0.967
8208	Knives and cutting blades for machines	8441	Machines making boxes, other paper products	0.966

borhood size to $\sqrt{N_i} \approx 34$ comparator industries and $\sqrt{N_l} \approx 11$ comparator locations (though our results are robust to varying the number of comparators: see Appendix A.3). These serve as our proxies for an industry-location's implied comparative advantage.

5 Main empirical results

We can now apply our approach to international and subnational datasets, which cover different countries, time periods and economic variables. We begin by constructing our similarity and density indices, and showing their explanatory power at the cross-section. Next, we study the growth rates of industry-location cells, which can only be defined for cells that start with a nonzero output. We then explore the extensive margin of growth by studying the appearance of industries that were not initially present in a particular location and also disappearance of industries with low implied comparative advantage values. For each analysis, we regress our measures of *implied* comparative advantage against current output levels, and then use the residuals to conduct out-of-sample regressions of either output growth or the appearance and disappearance of industries.

Figure 1: Distribution of similarity values.



5.1 Estimating *implied* comparative advantage

As argued above, density estimates an industry-location's comparative advantage, given the RpCA of its comparators. To see how well it fits, we estimate the following equation:

$$\log(R_{il,t_0}) = \alpha + \beta_I \log\left(\widehat{R}_{il}^{[I]}\right) + \beta_L \log\left(\widehat{R}_{il}^{[L]}\right) + \varepsilon_{il,t_0} \tag{5.1}$$

where ε_{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)	0.973***		0.774***
	(0.013)		(0.017)
Country Space density, 1995 (log)		0.990***	0.297***
		(0.039)	(0.019)
Observations	92,357	92,357	92,357
Adjusted R^2	0.636	0.493	0.654

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, since the values are using the same scale. 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 values than the CS density. When included in regressions together, both terms are still highly significant.

Table 5 shows the regressions for the US, India and Chile datasets. In all cases, both PS and CS density are significant ($p < 0.001$ in all cases); interestingly, they also yield coefficients that sum close to unity on average (from 0.75 to 1.1). As with the export data, R^2 values are substantial, especially for establishments. These results validate our first hypothesis: an industry-location's current comparative advantage can be estimated with some accuracy based on the values of similar industries and locations. However, as with any estimation, errors are made. Are these errors just noise, or do they carry information about the system's evolution? 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	0.516***	0.370***	0.508***	0.344***	0.165***
density (log)	(0.012)	(0.026)	(0.016)	(0.010)	(0.013)
Country Space	0.442***	0.401***	0.590***	0.510***	0.572***
density (log)	(0.008)	(0.026)	(0.013)	(0.011)	(0.010)
Observations	278,670	89,175	49,594	278,753	49,502
Adjusted R^2	0.275	0.368	0.403	0.795	0.697

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

5.2 Convergence of implied and observed comparative advantage

We have shown that our implied comparative advantage measures can explain much of the variation in the current comparative advantage patterns observed in the world. However, our 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 this by regressing the growth rate of the industry-location on the residuals from the first stage introduced in the previous section.[33] We use

[33] 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.

the standard definition of the annualized growth rate of y_{il}:

$$\dot{y}_{il} = \frac{1}{t_1 - t_0} \log \left(y_{il,t_1} / y_{il,t_0} \right) \tag{5.2}$$

where t_0 and t_1 are the initial year and final year, respectively. However, there are many industry-locations with an initial value of zero, for which we cannot define a growth rate. Likewise, cases in which the final value is zero are problematic because they introduce a hard boundary that would bias the estimates. We manage these issues by separately analyzing the intensive and extensive margins. Here we first examine the intensive margin by restricting our sample of industry-locations to those in which $y_{il,t_0} \neq 0$ and $y_{il,t_1} \neq 0$. In Section 5.3, we use a *probit* regression model to examine the probability of industry appearance (*i.e.*, growth from zero) and disappearance (*i.e.*, collapse into zero).

Our growth regression takes the following form, based on Equation 3.14:

$$\dot{y}_{il} = \alpha + \beta_\varepsilon \varepsilon_{il,t_0} + \gamma c_l + \delta d_i + e_{il}. \tag{5.3}$$

where ε_{il,t_0} is the residual term of the regression from the first stage. Note that while Equation 3.14 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 β_ε 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, and location and industry control variables. Finally, e_{il} is the error term.

Table 6 shows regressions of growth in international exports. The first three columns show that residuals from PS density, CS density, and *hybrid* density (the residual from the regression with both PS and CS density, *i.e.*, column 3 of Table 4) are highly significant predictors of industry-location growth ($p < 0.01$), and have the expected negative sign. Next, we include the initial global size of the industry and the location in question; these correspond with the industry-level and location-level components of the decomposition in Equation 3.2. 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.

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 them as the *ra-*

dial growth variables. Following the same standard compound growth formula as before, we calculate radial industry growth ($\dot{b}_i$) and radial location growth ($\dot{b}_l$) as:

$$\dot{b}_i = \frac{1}{t_1 - t_0} \log \left(\frac{\sum_l y_{il,t_1}}{\sum_l y_{il,t_0}} \right); \quad \dot{b}_l = \frac{1}{t_1 - t_0} \log \left(\frac{\sum_i y_{il,t_1}}{\sum_i y_{il,t_0}} \right) \tag{5.4}$$

These controls are an intuitive benchmark for our density indices, as they represent an alternative theory of growth (balanced growth). In fact, radial growth would account for all the variance in industry-location growth rates if all industries within a location grew at the same rate, or if all locations maintained their global market share in industries. Deviations from balanced growth thus mean that some industry-locations are increasing or decreasing their revealed comparative advantage. 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 maintain their economic and statistical significance, and increase the R^2 relative to column 6.

Our last specification captures all industry- and location-specific dynamics with fixed effects, subsuming the size and radial growth control variables as well as any other source of purely location-level of industry-level variation. Any additional explanatory power after controlling for these fixed effects (and initial industry-location size) must come entirely from industry-location interactions. Column 8 shows the baseline growth equation with location and industry fixed effects and initial location-industry size. Column 9 reintroduces the density residual and shows that its economic and statistical significance is undiminished. It is important to again point out that the density residual uses only base-year data and thus contains no information regarding future growth, while the coefficients on the fixed effects are calculated *ex post*. This means that our measures 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, we note that the robust and negative signs in the Columns 4-8 for initial industry-location exports mirror Rodrik (2013)'s observation of unconditional convergence at the industry level. But the significance of our density measures imply a richer structure in locations' convergence patterns.

Next, we apply the same process to our US, Chile and India datasets, over the maximum period available (Table 7). The density residuals are highly significant predictors of industry-location growth, both with and without controls ($p < 0.01$ for all cases). Put together, this suggests that our theoretical model – connecting *implied* comparative advantage to industry-location growth – is supported in a variety of contexts.

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 Space	-0.023***								
Density, 1995	(0.001)								
Residual, Country Space		-0.019***							
Density, 1995		(0.001)							
Residual, Hybrid			-0.024***		-0.024***		-0.019***		-0.024***
Density, 1995			(0.001)		(0.002)		(0.001)		(0.001)
Industry-location exports				-0.017***	0.000	-0.019***	-0.005***	-0.021***	-0.002**
1995 (log)				(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Location population				0.020***	0.000	0.024***	0.008***		
1995 (log)				(0.003)	(0.003)	(0.002)	(0.002)		
Global industry total				0.018***	0.002	0.020***	0.007***		
1995 (log)				(0.001)	(0.002)	(0.001)	(0.002)		
Mean location RpCA				0.013***	-0.010***	0.026***	0.007**		
1995 (log)				(0.003)	(0.003)	(0.002)	(0.003)		
Radial industry growth						0.993***	0.989***		
1995-2016						(0.018)	(0.018)		
Radial location growth						1.118***	1.015***		
1995-2016						(0.132)	(0.132)		
Observations	92,357	92,357	92,357	92,357	92,357	92,357	92,357	92,357	92,357
Adjusted R^2	0.150	0.153	0.161	0.137	0.185	0.299	0.328	0.412	0.439
Industry FE								Yes	Yes
Location FE								Yes	Yes

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

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, Hybrid	-0.045***	-0.042***	-0.031***	-0.026***	-0.048***	-0.050***	-0.055***	-0.044***	-0.320***	-0.177***
Density	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.005)	(0.008)
Initial Level		-0.004***		-0.004***		-0.004***		-0.009***		-0.172***
(log)		(0.001)		(0.000)		(0.001)		(0.002)		(0.007)
Initial Industry,		-0.003***		0.002***		-0.007***		0.008***		0.147***
Total (log)		(0.001)		(0.000)		(0.002)		(0.001)		(0.011)
Initial location,		0.002***		0.001***		-0.002*		0.014***		0.194***
Total (log)		(0.001)		(0.000)		(0.001)		(0.001)		(0.008)
Mean Location RpCA		0.004***		-		0.006***		-		0.140***
(log)		(0.001)				(0.001)				(0.010)
Radial industry		0.857***		0.754***		0.889***		0.698***		1.035***
growth (log)		(0.009)		(0.008)		(0.010)		(0.013)		(0.010)
Radial location		0.721***		0.628***		0.544***		0.486***		0.533***
growth (log)		(0.032)		(0.022)		(0.032)		(0.045)		(0.075)
Observations	278,670	278,670	278,753	278,753	89,175	89,175	49,502	49,502	49,594	49,594
Adjusted R^2	0.152	0.220	0.098	0.231	0.149	0.345	0.064	0.327	0.187	0.428

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

5.3 The extensive margin: Discrete appearances and disappearances

In previous sections, we analyzed the growth of industry-locations that already exist (*i.e.*, with non-zero exports, employment, establishment counts or payroll). In this section, we focus on the extensive margin, studying the *appearance* and *disappearance* of industry-locations. In particular, we expect that our measures of implied comparative advantage contain information on the overall fit of an industry in a location.

To do this, we first need to establish which industry-locations are likelier to be "present." The case is simple when using the US and Chilean data because they report the number of establishments. In these cases, an industry-location is present if at least one establishment is there. Formally, we capture this signal with the binary presence variable M_{il}:

$$M_{il,t_0} = \begin{cases} 1 & y_{il,t_0} \geq 1 \\ 0 & y_{il,t_0} = 0 \end{cases} \tag{5.5}$$

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. The simplest definition would be to define any nonzero export flow as an industry-location presence. In practice, however, trade data is full of one-time nonzero flows due to small re-exports, sales of used goods, or clerical errors; none of these would represent a present export industry in a meaningful sense. For this reason, we define an industry-location to be absent if $R_{il,t_0} < 0.05$, meaning that exports are less than 1/20th of the "expected" level. We consider an industry to be present if $R_{il} > 0.25$. As before, an appearance is a change from absent to present, and a disappearance is a change from present to absent. Thus, our definition of extensive margin change represents a fivefold relative increase or decrease.[34] See Tables 8 and 9 for the number and rates of presences, appearances and disappearances in each dataset.

We can now use our implied comparative advantage proxies to explain the appearance and disappearance of industries by location. First, a probit model estimates the probability of an industry-location presence based on PS and CS density:

$$P\left(M_{il} = 1\right) = \Phi\left(\alpha + \beta_I \log\left(\widehat{R}_{il}^{[I]} + c\right) + \beta_L \log\left(\widehat{R}_{il}^{[L]} + c\right)\right) \tag{5.6}$$

[34] While these thresholds are somewhat arbitrary, we obtain similar results using lower thresholds (a ceiling of 0.1 for absences and a floor of 0.1 for presences), or using the simple zero-nonzero definition.

where Φ is the Gaussian cumulative distribution function, and c is a constant (0.001) added to retain zero values.[35] As before, Equation 5.6 uses only information from t_0.

Going forward, we use the residual from 5.6 to quantify the gap between the world's current structure and its implied comparative advantage structure. As before, we hypothesize that there should be a significant association between this gap and future changes in structure (*i.e.*, appearances and disappearances). In particular, large *positive* residuals signify an "unexpectedly present" industry-location, and should thus predict disappearances, while large *negative* residuals signify an "unexpectedly absent" industry-location, and should thus predict appearances.[36] Finally, we include the total number of presences in a location (*"diversity"*) to test whether a Yule-like process is driving our appearance and disappearance results; in our context, this would mean we expect higher appearance rates from locations with higher diversity.[37]

In addition to the pseudo-R^2 statistic, we evaluate these predictions using the *area under the receiver-operating characteristic curve (AUC)*. The AUC is equivalent to the Mann-Whitney statistic, expressing the probability of ranking a true positive ahead of a false positive. 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 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, for the first year of each. Our PS and CS densities combined explain one half of the variance in industry-location; AUC values are also very high (close to 94% for hybrid models). Likewise, all coefficients are positive and highly significant, meaning that a high value for density is strongly indicative of an industry's presence.

Next, we use the residual term from these regressions to predict industry-location appearances and disappearances (Table 9). For all cases, the coefficients are highly significant, and have the expected sign. This means that *unexpectedly absent* industries tend to preferentially appear over time while *unexpectedly present* industries tend to disappear. This supports the hypothesis that industrial structure tends towards the deeper match between industries' requirements and locations' endowments (as captured by our proxies). Finally, we see diversity is significant, with the expected positive sign (or negative for disappearances); thus, a Yule process may also contribute to extensive margin dynamics.

[35] This constant is approximately equal to the 2^{nd} percentile for international trade-based density. We obtain the same results adding a constant of 0.01.

[36] These predictions are easier to comprehend if we compress and rearrange Equation 5.6 as $\varepsilon_{il} = M_{il} - \hat{M}_{il}$, where ε_{il} is the residual and $\hat{M}_{il}$ is the expected presence probability (based on density).

[37] In ecology, a Yule process means that each single organism has the same probability of giving birth, resulting in larger groups dominating future populations. Thanks to an anonymous referee for the idea.

Table 8: Probit regression of industry-location extensive margin, US, Chile and International

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	USA (establishments)			Chile (establishments)			International (exports)		
	Industry presences in 2003			Industry presences in 2005			Industry presences in 1995		
Product Space	1.269***		0.617***	0.942***		0.422***	1.321***		1.027***
density, initial year	(0.011)		(0.014)	(0.007)		(0.016)	(0.034)		(0.033)
Country Space		1.498***	1.150***		1.072***	0.891***		1.210***	0.427***
density, initial year		(0.010)	(0.014)		(0.015)	(0.016)		(0.053)	(0.030)
Observations		768,888			227,454			138,754	
Present industries		324622			55347			44085	
Presence rate		0.422			0.243			31.80%	
Area Under the Curve	0.883	0.925	0.934	0.856	0.929	0.936	0.932	0.898	0.936
Pseudo R^2	0.366	0.477	0.515	0.288	0.465	0.495	0.507	0.399	0.525

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

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)			Chile (establishments)			International (exports)		
	Industry appearances, 2003-11			Industry appearances, 2005-08			Industry appearances, 1995-2016		
Residual, Product	-1.694***			-1.780***			-0.685***		
Space density	(0.028)			(0.039)			(0.249)		
Residual, Country		-2.451***			-2.284***			-1.137***	
Space density		(0.020)			(0.033)			(0.198)	
Residual, Hybrid			-2.385***			-2.327***			-0.959***
Space Density			(0.022)			(0.034)			(0.220)
Location diversity,	0.048	1.190***	0.442***	0.457***	1.082***	0.660***	0.698***	0.729***	0.637***
initial year (log)	(0.030)	(0.029)	(0.029)	(0.029)	(0.031)	(0.029)	(0.180)	(0.147)	(0.168)
Initially absent		444,266			172,107			77,560	
Industry appearances		37681			11496			7063	
Appearance rate		0.0848			0.0668			0.0911	
Area Under the Curve	0.750	0.827	0.827	0.762	0.814	0.820	0.743	0.776	0.752
Pseudo R^2	0.0902	0.203	0.195	0.109	0.173	0.178	0.107	0.130	0.113
	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	USA (establishments)			Chile (establishments)			International (exports)		
	Industry disappearances, 03-11			Industry disappearances, 05-08			Industry disappearances, 1995-2010		
Residual, Product	2.992***			1.800***			1.125***		
Space density	(0.020)			(0.046)			(0.177)		
Residual, Country		2.524***			1.801***			1.258***	
Space density		(0.016)			(0.033)			(0.119)	
Residual, Hybrid			2.497***			1.757***			1.277***
Space Density			(0.019)			(0.033)			(0.157)
Location diversity,	0.988***	-1.018***	-0.305***	0.342***	-0.369***	-0.004	-0.680***	-0.924***	-0.618***
initial year (log)	(0.032)	(0.029)	(0.031)	(0.041)	(0.046)	(0.045)	(0.117)	(0.088)	(0.108)
Initially present		324,622			55,347			44,085	
Industry disappearances		45108			4762			3187	
Disappearance rate		0.139			0.086			0.0723	
Area Under the Curve	0.813	0.848	0.848	0.690	0.766	0.764	0.800	0.815	0.808
Pseudo R^2	0.192	0.235	0.229	0.0614	0.117	0.113	0.164	0.183	0.175

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

6 Robustness Exercises

6.1 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 implies that there is information in the least similar comparators as well.[38] The intuition is that if we can find industry or location pairs with a strong *negative* correlation, then a high comparative advantage in one would imply a low comparative advantage in the other.[39] Then as before, the residual between *observed* and *implied* comparative advantage should be negatively associated with growth.

Looking at the similarity distributions (Figure 1), we would expect this relationship to hold for location pairs: location similarities are distributed close to 0.5 (random), with nearly as many negatively correlated pairs (under 0.5) as positively correlated (over 0.5). For industries, however, there is an ominous lack of negatively correlated pairs. Going forward, we build density as before, but using the $\sqrt{N}$ *least* similar comparator locations and industries for each industry-location.

In stage 1 regressions (top half of Table 10), least-similar CS density performs as expected: its sign is negative, highly significant, and comparable in magnitude to that of the traditional CS density (though with a lower R^2). On the industry side, the signs are as expected for nearly all datasets. The exception is world trade (column 4), where least-similar PS density has a positive sign. This may be due to industries' tendency to follow a pattern of nestedness (Bustos et al. 2012) rather than traditional Ricardian specialization (Hidalgo and Hausmann 2009); see our discussion of Figure 1 above. Indeed, when we control for industry and location fixed effects (column 5), the intended negative coefficient appears. Finally, when raw establishment counts are used (columns 6 and 7), we also control for industry and location fixed effects to counteract the lack of RpCA normalization. With this correction, the coefficients are negative as expected.

Next, we use residuals from our stage 1 regressions to explain industry-location growth, as before. In bottom half of Table 10, the least-similar density residuals perform as expected: both CS and PS have negative signs, are comparable in magnitude to the original densities, and are highly significant ($p < 0.001$ in all cases).

[38] Thanks to two anonymous referees for suggesting this test.

[39] For example, the lowest similarity value we observe is between China and the US ($\phi_{China,USA} = 0.326$); this is far below random (0.5), meaning that we can expect one country to export what the other does not.

Table 10: Building density with least similar comparators.

	(1) USA employment	(2) USA payroll	(3) India employment	(4) Int'l exports	(5) Int'l exports	(6) USA estab.	(7) Chile estab.
	Cross-Section (Stage 1)						
Least-similar PS density, 1995 (log)	-0.591*** (0.011)	-0.194*** (0.010)	-0.126*** (0.021)	0.267** (0.105)	-0.949*** (0.039)	-0.156*** (0.005)	-0.268*** (0.015)
Least-similar CS density, 1995 (log)	-0.549*** (0.024)	-0.411*** (0.015)	-0.607*** (0.019)	-0.921*** (0.049)	-0.598*** (0.032)	-0.410*** (0.013)	-0.397*** (0.015)
Ind. & Loc. FEs	No	No	No	No	Yes	Yes	Yes
	Growth (Stage 2)						
Residual, PS density	-0.038*** (0.000)	-0.038*** (0.001)	-0.220*** (0.006)	-0.015*** (0.001)	-0.023*** (0.001)	-0.030*** (0.001)	-0.038*** (0.002)
Residual, CS density	-0.038*** (0.000)	-0.040*** (0.001)	-0.226*** (0.005)	-0.012*** (0.001)	-0.022*** (0.001)	-0.033*** (0.001)	-0.041*** (0.002)

Note: Each entry on this table represents the coefficient of a separate regression: the first row looking at PS density alone, and the second row looking at CS density alone. The residual from Stage 1 regressions (row 1 or 2) are used to predict the export growth in the same column of Stage 2 (row 3 or 4). Location-clustered robust standard errors in parentheses. Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6.2 Effect of time horizon

According to our theoretical model, the *implied* comparative advantage variable should be a better predictor of trends in *observed* comparative advantage over longer time-frames. In our growth regression models above, we chose the longest time interval possible for each dataset; Figure 2 shows adjusted R^2 values for international trade regressions (including base-year controls) over all possible year combinations. Each regression explains a sizable portion of the variation, with the lowest adjusted R^2 exceeding 8.5%, and a mean R^2 of 14.6%. Explanatory power generally improves as the interval increases (barring a possible continuity break between 1999 and 2000). This indicates that our measures capture a longer-term shift in economic structure, rather than a short-term mean reversion effect. This makes sense looking back at our theory: we had modelled the difference between true and observed comparative advantage as a shock that diminished over time; longer periods of time would thus yield more accurate predictions.

6.3 Testing the effect of the classification system

Could our results be an artifact of how industries are defined? If a classification system arbitrarily splits a single activity into two categories, then we would expect to see them in similar intensities. In the trade data classification, for example, we can see that HS6101

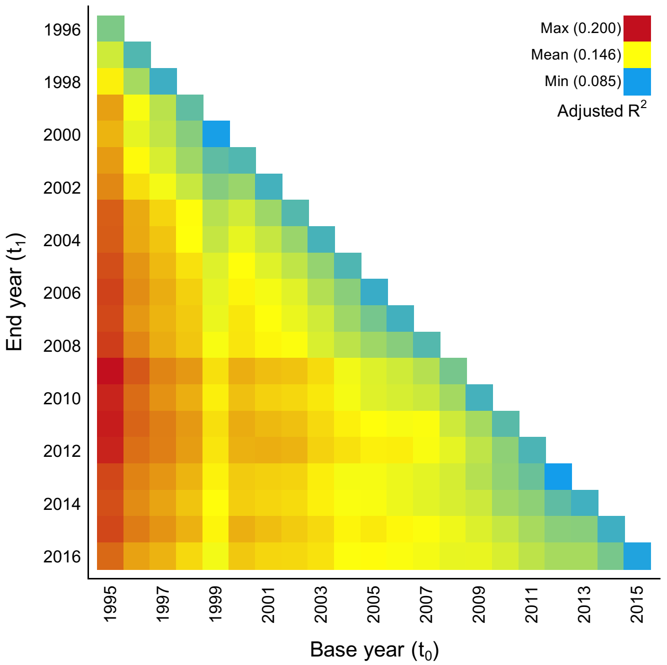


Figure 2: *Heat map of out of sample predictions of export growth, hybrid density model.*

contains "Men's overcoats" whereas HS6102 contains "Women's overcoats." In fact, in the top panel of Table 3, 9 out of 10 product pairs have the same 2-digit HS codes. To explore the possibility that our results are driven by such trivial cases, we calculate an adjusted PS density that excludes comparators from the same 2-digit HS categories.

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 essentially captures the information from classification system itself, since it relies on the 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.)

Table 11 present the results of these exercises, for the first and second stage regressions; we also repeat our main results for ease of comparison.[40] We can see that the results are essentially the same in these new specifications, with no major change in significance, sign, or explanatory power. Thus, our results are not artifacts of the classification system, and that our measure captures at least as much information as the classification provides.

6.4 Using other similarity matrices to build density

We have thus far shown evidence that the nature of an industry-location can be proxied using highly similar industries or locations – its *implied* comparative advantage. We are

[40] Panel (b) of Table 3 also gives the most similar industry pairs outside of the same HS chapter. These too appear quite related, despite coming from distinct sections of the classification.

somewhat agnostic to the exact functional form for our proxy: we do not claim that our similarity or density formulae are more accurate than other measures, but rather that ours can be derived from a Ricardian-inspired model of the world. We show that predictions from our theoretical model are also supported by other density measures.

Specifically, we test the original product space density proposed by Hausmann and Klinger (2006) and the taxonomy measure proposed by Zaccaria et al. (2014). HK density uses the binary presence-absence matrix to calculate a similarity measure based on co-production probabilities. Their density variable is calculated using all information in the matrix; the results to not substantially change when we limit the calculation to the top 34 most similar products. Taxonomy, 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 most important contributors to an industry. We again use the top 34 entries, to match the number used in our measure; however, the results are robust to using only the top-most entry. Finally, both measures typically base their presence-absence matrix on RCA; here we report versions calculated using RpCA (to match our own), though results are virtually unchanged when using RCA instead. We expect these measures to perform similarly to ours. However, their use of the binary presence-absence matrix rather than continuous RpCA values may result in a loss of explanatory power. We also no longer expect the coefficient of HK density to approach unity, since its scale is not the same as the dependent variable; the coefficient should still be positive, however.

Table 11 shows the results for both stages. As expected, all first-stage terms are positive and highly significant. We also find little effect of these adjustments in the second stage: the residual term is always negative (as expected), highly significant, and with a similar magnitude to that of our original density formula. This appears to reject the possibility that our findings are an artifact of the exact similarity or density formula used.

7 Conclusions

In this paper we present and test a model showing that the observed comparative advantage of an industry in a location follows a pattern that can be discerned from related industries in that same location (product space density) or the same industry in related locations (country space density). Moreover, we present evidence supporting our model's prediction that the error term in the predicted pattern is not pure noise, but instead carries information regarding the future trend of that industry-location. 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,

Table 11: Building densities from other industry similarity matrices.

	Stage 1		Stage 2	
	coefficient	R^2	coefficient	R^2
Product Space density (log)	0.975*** (0.013)	0.636	-0.018*** (0.001)	0.325
PS density, excluding same 2-digit industry (log)	0.956*** (0.014)	0.617	-0.018*** (0.001)	0.320
Mean RpCA of others in same 2-digit category (log)	0.933*** (0.006)	0.610	-0.014*** (0.001)	0.322
Taxonomy (log)	1.071*** (0.023)	0.476	-0.010*** (0.001)	0.308
PS Density, HK method (log)	1.815*** (0.063)	0.485	-0.019*** (0.003)	0.318

Note: Each entry represents a separate regression, over a consistent sample (N=91,828). The columns associated with Stage 1 replicate Table 4 with different PS densities; each row tests a different density specification using only 1995 data. The residual from Stage 1 regressions, along with controls in Table 6 are used to predict the export growth in Stage 2; coefficients from controls not reported. Country-clustered robust standard errors in parentheses. Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

as well as subnational data for the US, India and Chile. We show evidence that they operate at the intensive and extensive margins, that they can be produced using *least* similar (negatively correlated) comparators, that they are not artifacts of the classification system used, and that they operate most intensely at longer time horizons.

An important question is why would our measures carry information about the growth of an industry-location, even after controlling for location and industry effects (including their overall growth rates). The interpretation we model is that each industry-location is affected by a shock that causes a deviation of its output from their ”true” comparative advantage levels. In this interpretation, since over time the expected value of shock is zero, the underlying fundamentals are increasingly expressed over time, and are predicted by our measures. An alternative interpretation is that the similarity measures we use are capturing 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 is affected by the presence of related industries through spillovers. The fact that these take time would explain why our predictive power peaks at periods of a decade or more. From a Ricardian viewpoint, on the other hand, the conjecture would be that mastery of specific technologies affects the productivity of related industries, a feature not incorporated into current Ricardian models. Efforts to improve one industry's productivity may spill over into related industries. Unexploited aspects of technological relatedness are reflected in the difference between a country's structure and the international norm. These differences diminish over time as firms exploit technological spillovers. Future research could test these rival hypotheses.

One salient feature of the similarity metrics we use is their symmetry. In reality, we expect a directionality among industries: as the countries climb the development ladder, they move into more sophisticated products (Hausmann et al. 2014). For example, we might observe that all countries exporting pharmaceuticals also export chemical products, but not all chemical exporters are pharmaceutical exporters; in other words, one industry could be a prerequisite for another. This idea is present in Hirschman (1958)'s seminal work, where he differentiates the importance of backward versus forward linkages in a development strategy.[41] Future studies could incorporate this asymmetry.

The results in our paper may be important for policymakers developing strategies for growth. First, they allow policymakers to estimate potential comparative advantage, even for industries currently absent from their area. Second, they can help identify over- or under-performing industries depending on other industries present in the location; this in turn may help estimate the growth potential of local industries. Finally, our model might also help identify pervasive failures and obstacles, using historical data to identify which industries had the predicted potential to grow and why this potential was not realized.[42] These issues might then be addressed, if policymakers so choose.

[41] See Maurseth and Verspagen (2002) and O'Clery et al. (2018) for recent uses of asymmetric metrics.

[42] For example, Japan's largest negative residual is for HS 9306: "Munitions of war." The interpretation is that Japan's weapons export industry is "unexpectedly absent," given its export of similar products. This makes intuitive sense, since Japan's reasons for not producing those goods are historical, not technological.

References

Asher, Sam, and Paul Novosad. 2017. “Politics and local economic growth: Evidence from India.” *American Economic Journal: Applied Economics* 9 (1): 229–73.

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

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

Beaudry, Catherine, and Andrea Schiffauerova. 2009. “Who’s right, Marshall or Jacobs? The localization versus urbanization debate.” *Research Policy* 38 (2): 318–337.

Boschma, Ron. 2017. “Relatedness as driver of regional diversification: A research agenda.” *Regional Studies* 51 (3): 351–364.

Boschma, Ron, and Gianluca Capone. 2015. “Institutions and diversification: Related versus unrelated diversification in a varieties of capitalism framework.” *Research Policy* 44 (10): 1902–1914.

Boschma, Ron, Gaston Heimeriks, and Pierre-Alexandre Balland. 2014. “Scientific knowledge dynamics and relatedness in biotech cities.” *Research Policy* 43 (1): 107–114.

Boschma, Ron, Asier Minondo, and Mikel Navarro. 2012. “Related variety and regional growth in Spain.” *Papers in Regional Science* 91 (2): 241–256.

———. 2013. “The Emergence of New Industries at the Regional Level in Spain: A Proximity Approach Based on Product Relatedness.” *Economic Geography* 89 (1): 29–51.

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

Bustos, Sebastián, and Muhammed A Yildirim. 2019. *Production Ability and Economic Growth.* Technical report. Center for International Development at Harvard University Fellows Working Paper No:110.

Caldarelli, Guido, Matthieu Cristelli, Andrea Gabrielli, Luciano Pietronero, Antonio Scala, and Andrea Tacchella. 2012. “A network analysis of countries? export flows: firm grounds for the building blocks of the economy.” *PloS one* 7 (10): e47278.

Caliendo, Lorenzo, Fernando Parro, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte. 2017. "The impact of regional and sectoral productivity changes on the US economy." *The Review of Economic Studies* 85 (4): 2042–2096.

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

Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. "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* 124 (1): 205–248.

Cristelli, Matthieu, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. 2013. "Measuring the intangibles: A metrics for the economic complexity of countries and products." *PloS one* 8 (8): e70726.

Davis, Donald R, and Jonathan I Dingel. 2014. *The comparative advantage of cities.* Technical report. National Bureau of Economic Research.

Deardorff, Alan V. 1984. "Testing trade theories and predicting trade flows." *Handbook of International Economics* 1:467–517.

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

———. 2015. "Defining clusters of related industries." *Journal of Economic Geography* 16 (1): 1–38.

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

Duda, Richard O, Peter E Hart, and David G Stork. 2012. *Pattern classification.* John Wiley & Sons.

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

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

Ellison, Glenn, Edward L Glaeser, and William R Kerr. 2010. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." *American Economic Review* 100 (3): 1195–1213.

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

Hanlon, W Walker, and Antonio Miscio. 2017. "Agglomeration: A long-run panel data approach." *Journal of Urban Economics* 99:1–14.

Hausmann, Ricardo, César A Hidalgo, Sebastián Bustos, Michele Coscia, Alexander Simoes, and Muhammed A. Yıldırım. 2014. *The Atlas of Economic Complexity: Mapping Paths to Prosperity.* The MIT Press.

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

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

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

Hidalgo, César A, Pierre-Alexandre Balland, Ron Boschma, Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, Canfei He, Dieter F Kogler, Andrea Morrison, et al. 2018. "The principle of relatedness." In *International Conference on Complex Systems,* 451–457. Springer.

Hidalgo, César A, and Ricardo Hausmann. 2009. "The building blocks of economic complexity." *Proceedings of the National Academy of Sciences of the United States of America* 106 (26): 10570–10575.

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

Hirschman, Albert O. 1958. *The strategy of economic development.* New Haven, Connecticut: Yale University Press.

Liang, Jiaochen, and Stephan J Goetz. 2018. "Technology intensity and agglomeration economies." *Research Policy* 47 (10): 1990–1995.

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

Lu, Ren, Min Ruan, and Torger Reve. 2016. "Cluster and co-located cluster effects: An empirical study of six Chinese city regions." *Research Policy* 45 (10): 1984–1995.

Marshall, Alfred. 1890. *Principles of economics.* Macmillan / Company.

Maurseth, Per Botolf, and Bart Verspagen. 2002. "Knowledge spillovers in Europe: a patent citations analysis." *Scandinavian Journal of Economics* 104 (4): 531–545.

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

O'Clery, Neave, Muhammed A Yildirim, and Ricardo Hausmann. 2018. "Productive Ecosystems and the Arrow of Development." ArXiv preprint arXiv:1807.03374.

Petralia, Sergio, Pierre-Alexandre Balland, and Andrea Morrison. 2017. "Climbing the ladder of technological development." *Research Policy* 46 (5): 956–969.

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

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

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

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

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

Tacchella, Andrea, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. 2012. "A new metrics for countries' fitness and products' complexity." *Scientific reports* 2:723.

Tacchella, Andrea, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. 2013. "Economic complexity: conceptual grounding of a new metrics for global competitiveness." *Journal of Economic Dynamics and Control* 37 (8): 1683–1691.

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

Zaccaria, Andrea, Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. 2014. "How the taxonomy of products drives the economic development of countries." *PloS one* 9 (12): e113770.

A Appendix

A.1 Connecting Similarity Coefficients To Underlying Requirements and Endowments

A.1.1 Proof of general case with n-dimensional sphere and 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 ψ_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 λ_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, ψ_i, and the endowments of the location, λ_l. More concretely,

$$r_{il} = f\left(d(\psi_i, \lambda_l)\right) \tag{A.1}$$

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 ψ_i and λ_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:

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

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 ψ_i and λ_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:

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}]}{E[(r_{il} - \bar{r}_i)^2|\psi_i]} \tag{A.3}$$

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.1, we can calculate:

$$\begin{aligned} E[r_{il}^2|\psi_i] &= \int\limits_{\lambda_l \in \mathbb{S}} \left(f(d(\psi_i, \lambda_l))\right)^2 \mathrm{d}\lambda_l \\ E[r_{il}|\psi_i] &= \int\limits_{\lambda_l \in \mathbb{S}} f(d(\psi_i, \lambda_l))\mathrm{d}\lambda_l \end{aligned}$$

Similarly, we can write the numerator in A.3 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:

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

Using Equation A.3, 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 ψ_i and $\psi_{i'}$. For the sake of brevity, we will call it σ^2. We can write the numerator as:

$$\begin{aligned} E[r_{il}^2|\psi_i] - E[r_{il}r_{i'l}|\psi_i, \psi_{i'}] &= \int\limits_{\lambda_l \in \mathbb{S}} \left(f(d(\psi_i, \lambda_l))\right)^2 \mathrm{d}\lambda_l - \int\limits_{\lambda_l \in \mathbb{S}} f(d(\psi_i, \lambda_l))f(d(\psi_{i'}, \lambda_l))\mathrm{d}\lambda_l \\ &= \int\limits_{\lambda_l \in \mathbb{S}} f(d(\psi_i, \lambda_l)) \left[f(d(\psi_i, \lambda_l)) - f(d(\psi_{i'}, \lambda_l))\right] \mathrm{d}\lambda_l \end{aligned}$$

Let's move the origin to ψ_i and define $\psi_{i'} = \psi_i + \Delta_{ii'}$. Hence, the expression above can be written as a function of $\Delta_{ii'}$ as:[43]

$$g(\Delta_{ii'}) = \int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \left[f(d(0,\lambda_l)) - f(d(\Delta_{ii'},\lambda_l)) \right] \mathrm{d}\lambda_l$$

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

$$g(\Delta_{ii'} + \delta_{ii'}) = \int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \left[f(d(0,\lambda_l)) - f(d(\Delta_{ii'} + \delta_{ii'},\lambda_l)) \right] \mathrm{d}\lambda_l$$

Hence:

$$g(\Delta_{ii'} + \delta_{ii'}) - g(\Delta_{ii'}) = \int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \left[f(d(\Delta_{ii'} + \delta_{ii'},\lambda_l)) - f(d(\Delta_{ii'},\lambda_l)) \right] \mathrm{d}\lambda_l$$

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

$$\frac{g(\Delta_{ii'} + \delta_{ii'}) - g(\Delta_{ii'})}{\delta_{ii'}} = \int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \frac{f(d(\Delta_{ii'} + \delta_{ii'},\lambda_l)) - f(d(\Delta_{ii'},\lambda_l))}{\delta_{ii'}} \mathrm{d}\lambda_l$$

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:

$$\begin{aligned} \frac{\mathrm{d}g(\Delta_{ii'})}{\mathrm{d}\Delta_{ii'}} &= \int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \frac{f(d(\Delta_{ii'},\lambda_l - \delta_{ii'})) - f(d(\Delta_{ii'},\lambda_l))}{\delta_{ii'}} \mathrm{d}\lambda_l \\ &= -\int\limits_{\lambda_l \in \mathbb{S}} f(d(0,\lambda_l)) \frac{\mathrm{d}f(d(\Delta_{ii'},\lambda_l))}{\mathrm{d}\lambda_l} \mathrm{d}\lambda_l \end{aligned}$$

$d(\Delta_{ii'},\lambda_l)$ increase as λ_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:

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

[43] Note that, for the one-dimensional circle with circumference 1, this function is:

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

Following Equation A.1, 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 ψ_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.1.2 Proof of simple case on unit sphere and a specified functional form.

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 ψ_i and λ_l parameters on a circle of perimeter 1. We can define the similarity $\phi_{ii'}$ between two industries i and i' given by:

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

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

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

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

$$\text{corr}\{r_i, r_{i'}\} = \frac{E[(r_{il} - \bar{r}_i)(r_{i'l} - \bar{r}_{i'})|\psi_i, \psi_{i'}]}{E[(r_{il} - \bar{r}_i)^2|\psi_i]} \tag{A.4}$$

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.4 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.4 becomes:

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

We can write the numerator in Equation A.4 as:

$$\begin{aligned} E[(\tilde{r}_{il} - \bar{\tilde{r}}_i)(\tilde{r}_{i'l} - \bar{\tilde{r}}_{i'})|\psi_i, \psi_{i'}] &= \int_0^1 \left(d^2(\psi_i, \lambda_l) - \frac{1}{12}\right)\left(d^2(\psi_{i'}, \lambda_l) - \frac{1}{12}\right) d\lambda_l \\ &= \int_0^1 [d(\psi_i, \lambda_l) d(\psi_{i'}, \lambda_l)]^2 d\lambda_l - \frac{1}{144} \end{aligned} \tag{A.5}$$

To calculate the integral, we will measure all the distances on the circle relative to ψ_i. Let's define $\Delta_{ii'} \equiv d(\psi_i, \psi_{i'})$. We can write the integral in Equation A.5 as:

$$\begin{aligned} \int_0^1 [d(\psi_i, \lambda_l) d(\psi_{i'}, \lambda_l)]^2 d\lambda_l = & \int_0^{1/2} [y(y - \Delta_{ii'})]^2 dy \\ & + \int_{1/2}^{1/2+\Delta_{ii'}} [(1-y)(y - \Delta_{ii'})]^2 dy \\ & + \int_{1/2+\Delta_{ii'}}^{1} [(1-y)(1 - y + \Delta_{ii'})]^2 dy \end{aligned} \tag{A.6}$$

The first integral in Equation A.6 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.6 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.6 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.4, we obtain:

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

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

$$\phi_{ii'} = (1 + \text{corr}\{r_i, r_{i'}\})/2 = 1 - 15\left(d(\psi_i, \psi_{i'}) - d^2(\psi_i, \psi_{i'})\right)^2 \tag{A.7}$$

A.1.3 Simulating the similarity distribution with noise.

We tested the validity of expression in Equation A.7 for different noise-to-single 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

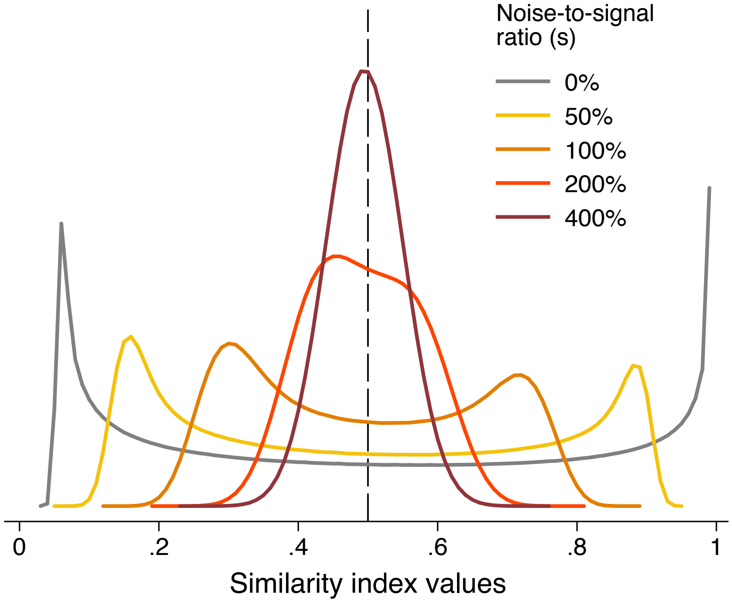


Figure A.1: *Simulated similarity distributions for different noise levels.*

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 Additional Empirical Results

A.2.1 Additional tables for empirical results described in main text.

A.2.2 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 A.1 applies this process to our international trade dataset. We find that the explanatory power of our out-of-sample hybrid model is comparable to that of the in-sample model (R^2 values are 62.2% and 55.7% for regressions of current export levels, and 18.7% versus 18.5% for regressions of export growth). Furthermore, adding the in-sample density terms to the out-of-sample dataset yields a negligible marginal contribution to R^2. Finally, combining the in-sample and out-of-sample predictors shows a marginally higher R^2 but with drastically reduced significance, indicating a high degree of co-linearity between the two types of variables. This suggests that endogeneity is not driving our results.

Table A.1: Cross-validating OLS regression of international exports and export growth by industry-location.

	(1)	(2)	(3)	(4)	(5)	(6)
	Observed Comparative Advantage			Growth in exports (log)		
	RpCA of Exports (log), 1995					
Product Space density	0.916***		0.940***			
out-of-sample, 1995 (log)	(0.025)		(0.065)			
Country Space density	0.150***		0.063			
out-of-sample, 1995 (log)	(0.038)		(0.046)			
Product Space density		0.830***	-0.035			
in-sample, 1995 (log)		(0.029)	(0.066)			
Country Space density		0.357***	0.121**			
in-sample, 1995 (log)		(0.049)	(0.053)			
Residual Product Space density				-0.012***		-0.006***
out-of-sample, 1995				(0.002)		(0.002)
Residual Country Space density				-0.014***		-0.009**
out-of-sample, 1995				(0.002)		(0.004)
Residual Product Space density					-0.012***	-0.006***
in-sample, 1995					(0.002)	(0.002)
Residual Country Space density					-0.014***	-0.006
in-sample, 1995					(0.002)	(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 *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

A.2.3 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 regressions 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 A.2 shows the results. For product subsamples (based on the HS chapters), it

Table A.2: Cross-sectional Regressions for Product and Country Groups.

Products		Countries	
HS Chapter	Adjusted R^2	By Income	Adjusted R^2
Plastics & Rubbers	0.231	Upper middle income	0.212
Processed Metals	0.201	Low income	0.201
Electronics, Machinery & Equipment	0.200	Lower middle income	0.186
		High income	0.137
Automotive, Planes, Ships & Related	0.193		
Agricultural Products	0.188	By region	Adjusted R^2
Medical, Consumer & Other	0.186	Europe & Central Asia	0.211
Chemicals & Related	0.185	South Asia	0.187
Processed Foodstuffs	0.179	Middle East & North Africa	0.183
Wood Products	0.171	East Asia & Pacific	0.180
Processed Stone & Glass	0.168	Sub-Saharan Africa	0.169
Extractives	0.152	Americas	0.167
Apparel & Textiles	0.140		

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.

A.2.4 Using RCA

For each dataset, we build the similarity and density indices for measuring implied comparative advantage, as described above. Our first step is to normalize the export, employment and payroll data to focus on the observed comparative advantage of each industry-location, and to facilitate comparison across location, industry and time. In most cases, we use Balassa's revealed comparative advantage (RCA) index (Balassa 1964) of location

l in industry i in year t_0:

$$R_{il,t_0} = \frac{y_{il,t_0} / \sum_i y_{il,t_0}}{\sum_l y_{il,t_0} / \sum_l \sum_i y_{il,t_0}} \tag{A.8}$$

where y_{il,t_0} is the export, employment or payroll value. We do not normalize the number of establishments. Note that for some industries, these R values can get quite high (in the thousands); these rare cases can have an outsize impact on our similarity correlations. For this reason, we cap RCA at 5 and establishments at 100 when building our similarity indices (Equations 3.4 and 3.5 above); these caps correspond with the 97th to 99th percentile, depending on the dataset used.[44] These caps only directly affect the similarity indices; the density index still uses uncapped RCA or establishments (weighted by the similarity indices).

Table A.3 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. However, compared to RpCA in Table 4, we observe lower R^2 values overall.

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

	(1)	(2)	(3)
	Observed Comparative Advantage RCA of Exports (log), 1995		
Product Space density, 1995 (log)	0.981***		0.826***
	(0.020)		(0.021)
Country Space density, 1995 (log)		0.741***	0.231***
		(0.023)	(0.014)
Observations	92,355	92,355	92,355
Adjusted R-squared	0.420	0.260	0.435

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

Table A.4 shows a set of growth regressions using our international export data. The

[44] Our main results are robust to using uncapped RCA, or using logs (plus a constant) instead of a cap.

dependent variable is the growth rate in the industry-location cell. The first three columns in Table A.4 use as independent variable the error terms from the three regressions in Table A.3. They show that the residual using both product space and country space densities, as well as both of them combined are highly significant predictors of growth and explain between 15 and 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. Column 4 shows that these variables, on their own, are significantly related to subsequent growth; 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. 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. 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.

Finally, in Table A.5 we show the results with density variables built with the least similar industries or locations. Here, all the coefficients are negative as expected.

Table A.4: 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	-0.024***		-0.016***		-0.016***		-0.014***		-0.016***
Space Density, 1995	(0.001)		(0.002)		(0.002)		(0.001)		(0.001)
Residual, Country		-0.020***	-0.008***		-0.005**		-0.006***		-0.009***
Space Density, 1995		(0.001)	(0.002)		(0.002)		(0.001)		(0.001)
Industry-location				-0.017***	-0.003**	-0.018***	-0.005***	-0.021***	-0.003***
exports, 1995 (log)				(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Location total				0.016***	0.000	0.025***	0.010***		
1995 (log)				(0.002)	(0.003)	(0.002)	(0.002)		
Industry total				0.018***	0.006***	0.020***	0.008***		
1995 (log)				(0.001)	(0.002)	(0.001)	(0.001)		
Mean location RCA				0.023***	0.010	0.017***	0.006		
1995 (log)				(0.007)	(0.007)	(0.005)	(0.005)		
Radial industry growth						0.991***	0.998***		
1995-2016						(0.018)	(0.018)		
Radial location growth						1.116***	1.087***		
1995-2016						(0.113)	(0.110)		
Observations	92,355	92,355	92,355	92,355	92,355	92,355	92,355	92,355	92,355
Adjusted R-squared	0.148	0.131	0.155	0.133	0.165	0.300	0.328	0.412	0.440
Industry FE								Yes	Yes
Location FE								Yes	Yes

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

Table A.5: Regressions with Least Similar Industries.

	USA employment	USA payroll	India employment	Int'l exports
	Cross-Section (Stage 1)			
PS Density, (log)	-0.601***	-0.274***	-0.646***	-0.902***
	(0.010)	(0.005)	(0.073)	(0.038)
CS Density (log)	-0.535***	-0.123***	-0.549***	-0.644***
	(0.027)	(0.014)	(0.017)	(0.048)
	Growth (Stage 2)			
Residual, PS Density	-0.037***	-0.034***	-0.255***	-0.016***
	(0.000)	(0.001)	(0.008)	(0.001)
Residual, CS Density	-0.037***	-0.031***	-0.250***	-0.017***
	(0.000)	(0.001)	(0.005)	(0.001)

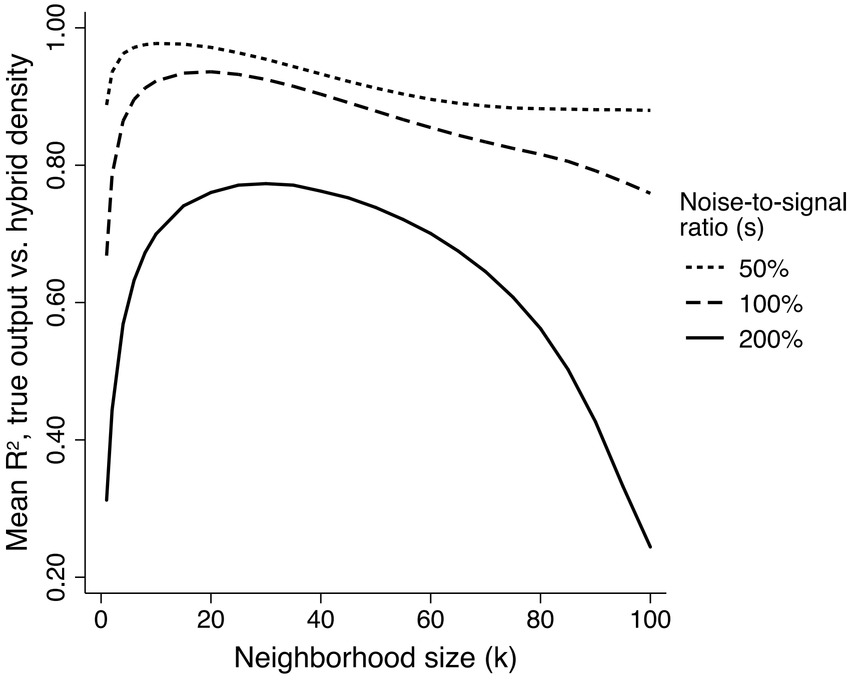
Note: We calculate the PS and CS densities using 34 least similar industries and 11 least similar countries, respectively, instead of most similar ones. Each entry on this table represents the coefficient of a separate regression with the corresponding PS density. In rows associated with Stage 1 replicates Table 4 with different Product Space densities. The residual from Stage 1 regressions are used to predict the export growth in Stage 2. Location-clustered robust standard errors in parentheses. Significance given as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.3 Modeling and Testing the Optimal Number of Comparators

In our Theory section, presented 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 10 most similar locations to its 50 most similar locations – the additional comparators will become increasingly dissimilar from the original location, diminishing the accuracy of the *implied* comparative advantage measure. Thus, we would expect to see an inverted U-shaped relationship between the number of comparators used (k) and the accuracy of our weighted average.

We can explore this expectation using our simulation. Normally, the simulation had fixed k (the number of comparators used to build the density variable) equal to $\sqrt{N}$ (recall that we had set $N = 100$). We now relax this constraint, in order to see what happens as k varies. Figure A.2 presents the results, across three levels of noise. As expected, we find

Figure A.2: Observed vs. Implied Comparative Advantage for Different Noise Levels

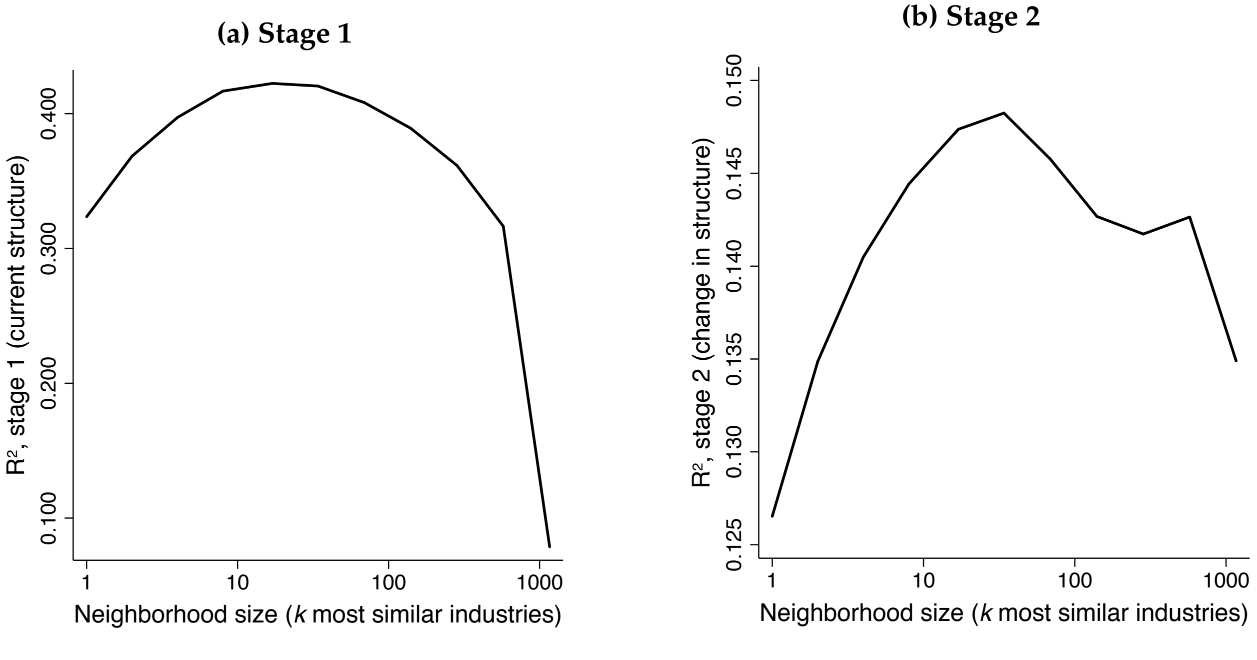


an inverted-U shaped relationship between k and the correlation between *true* and *implied* comparative advantage: the correlation grows stronger as we include more comparators, then quickly peaks around the lower-middle range, k = 8 to 35 (depending on the level of noise). After that, adding more comparators 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 small changes in k, as long as we stay in that lower-middle range (including our default $k = \sqrt{N}$ setting).

> **Hypothesis A1:** 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.

Does this hypothesis bear out empirically? In the international trade data regressions above, we constructed the PS and CS density measures using the top 34 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.

Figure A.3: Changing Number of Comparators.



(a) Stage 1

(b) Stage 2

Figures A.3a and A.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 A.2 – see especially the *high noise* results): the R^2 values peak near the geometric midpoint (at k=34 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).

B Motivation Based on Factor Content of Production

Our Ricardian-inspired model can essentially 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. Here, 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 comparative advantage of a location in an industry can be inferred from its comparative advantage in industries that have similar factor requirements or locations that have similar factor endowments. Interestingly, our results can be derived without information regarding production functions or factor endowments.

B.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 Hecksher-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. 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.

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.

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:

$$AY = F \tag{B.1}$$

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 B.1 by the B matrix, we obtain:

$$Y = BF \tag{B.2}$$

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.

B.2 Similarities between the factor requirements of two industries

We will now derive a measure of input similarity of two industries, using Equation B.2. 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:

$$r_{il} = \sum_f B_{if} F_{fl} \tag{B.3}$$

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:

$$r_i = B_i F \tag{B.4}$$

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 B.4, average production of industry i can be calculated as:

$$\overline{r}_i = \frac{\sum_l r_{il}}{N_l} = \sum_f B_{if} \frac{\sum_l F_{fl}}{N_l} = \sum_f B_{if} \overline{F}_f \tag{B.5}$$

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

$$r_i - \overline{r}_i = B_i (F - \overline{F}) \tag{B.6}$$

where $\overline{F}$ is a $N_f \times N_l$ matrix that repeats in each row f the average endowment of the world in that factor $\overline{F}_f$. Using Equation B.6, we can relate the observed covariance of the rows of the Y matrix to those of the unobserved B matrix:

$$(r_i - \overline{r}_i)(r_{i'} - \overline{r}_{i'})^t = B_i (F - \overline{F})(F - \overline{F})^t B_{i'}^t \tag{B.7}$$

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

$$C = U \Sigma U^t \tag{B.8}$$

where U is a unitary matrix formed by the eigenvectors of C and Σ is a diagonal matrix whose elements are eigenvalues of C. If we define $\tilde{B}_i = B_i U$, then we can write the right hand side of Equation B.7 as:

$$(r_i - \overline{r}_i)(r_{i'} - \overline{r}_{i'})^t = \sum_f \tilde{B}_{if} \tilde{B}_{i'f}^t \sigma_f \tag{B.9}$$

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

$$\sum_f \tilde{B}_{if} \tilde{B}_{i'f}^t \sigma_f = \sigma \tilde{B}_i \tilde{B}_{i'}^t = \sigma B_i U U^t B_{i'}^t = \sigma B_i B_{i'}^t \tag{B.10}$$

Dividing both sides of Equation B.10 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:

$$\mathrm{corr}\{r_i, r_{i'}\} = \frac{(r_i - \bar{r}_i)(r_{i'} - \bar{r}_{i'})^t}{\sigma_{r_i}\sigma_{r_{i'}}} \approx \frac{\sigma}{\sigma_{r_i}\sigma_{r_{i'}}} B_i B_{i'}^t \tag{B.11}$$

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

$$\phi_{ii'} = (1 + \mathrm{corr}\{r_i, r_{i'}\})/2 \tag{B.12}$$

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

B.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^{th} column vectors of Y and F matrices respectively, the output of a location is related to its factor endowments by:

$$r_l = BF_l \tag{B.13}$$

Note that our calculations in Section 2.1.1 can be replicated here because if we take the transposes of both sides in Equation B.13, we will arrive to an expression similar to Equation B.4. Assuming that the columns of B matrix are independently and identically distributed, we can write (akin to Equation B.11):

$$\mathrm{corr}\{r_l, r_{l'}\} = \frac{(r_l - \bar{r}_l)^t(r_{l'} - \bar{r}_{l'})}{\sigma_{r_l}\sigma_{r_{l'}}} \approx \frac{\sigma'}{\sigma_{r_l}\sigma_{r_{l'}}} F_l^t F_{l'} \tag{B.14}$$

where $\bar{r}_l$ is the average production of location l, σ_{r_l} is the standard deviation of r_l, σ' is the diagonal of the covariance matrix ($(B - \overline{B})^t(B - \overline{B}) \approx \sigma' I_{N_f}$). We renormalize the correlation to build a similarity metric between 0 and 1 by adding 1 and dividing by 2. Hence, we can estimate a measure of the similarity between the factor endowments of two locations, l and l' as:

$$\phi_{ll'} = (1 + \mathrm{corr}\{r_l, r_{l'}\})/2 \tag{B.15}$$

where 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.

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

$$\widehat{A}\widehat{r} = \widehat{F} \tag{B.16}$$

All the previous results will follow in this re-normalized 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:

$$r_{il} = X_{il} + C_{il} \tag{B.17}$$

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:

$$r_i - \overline{r}_i = (X_i - \overline{X}_i) + (C_i - \overline{C}_i) \tag{B.18}$$

Assuming homothetic preferences worldwide, and normalizing each industry element by its size, we can assume that $C_i = \overline{C}_i$. Therefore, correlations of columns of Y can be inferred from correlations of columns of X. Similarly, we can also look at the column vectors of Y and X:

$$r_l - \overline{r}_l = (X_l - \overline{X}_l) + (C_l - \overline{C}_l) \tag{B.19}$$

Again, assuming homothetic preferences worldwide, and normalizing each location by its size then each country would consume the same share of products, implying that $C_l = \overline{C}_l$.

Consequently, correlations between the columns of Y can be inferred from the correlations between the columns of X.

B.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 B.11. We randomly select 5,000 A^{-1} and F matrices and test the validity of this equation. We note that the random selection of both matrices simultaneously puts no inherent structure into these matrices and in reality we expect to observe more structures matrices. Even in the random case, the correlation between the actual and estimated numbers exhibit is 0.532 ∓ 0.014. We also test whether the correlation between pairs of columns of Y is related to the correlation between the corresponding columns of factor endowments F as suggested by Equation B.14 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 [0.7150.853] across all simulations). However, we need not fix the neighborhood filter at $k = 50$. In Figure 4, the uppermost line shows the effect of neighborhood size on the R^2. We see that the highest R^2 value is found at $k = 4$.

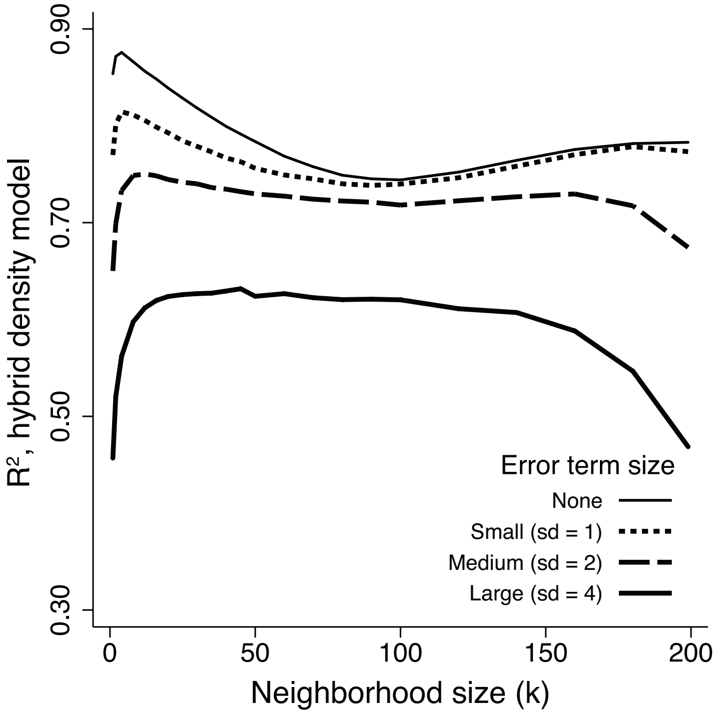


Figure 4: *Simulation of association between underlying output and hybrid density model, by size of neighborhood and noise level.*

Finally, we can extend our simulation to examine the effect of noise in the observed output. Beginning with the $Y = A^{-1}F$ used above, suppose that observed output, $\tilde{r}_{il}$, is affected by a random error term, ε_{il}, with a normal distribution around a mean of zero:

$$\tilde{r}_{il} = r_{il} + \varepsilon_{il} \tag{B.20}$$

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 the standard deviation of r_{il}, respectively.

In Figure 4, 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.

B.6 Conclusions related to HOV

Ricardian models are reduced-form models, where other elements are subsumed in the labor productivity parameters. Here, 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.

References for Appendix B

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

———. 1995. “The Case of the Missing Trade and Other Mysteries.” *American Economic Review* 85 (5): 1029–1046.

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

———. 2010. “The structure of factor content predictions.” *Journal of International Economics* 82 (2): 195–207.

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

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