Growth Lab

From Products to Capabilities: Constructing a Genotypic Product Space

Ulrich Schetter, Dario Diodato, Eric Protzer, Frank Neffke, and Ricardo Hausmann

Growth Lab Working Paper Series
No. 230

June
2024

GROWTH LAB
HARVARD KENNEDY SCHOOL
79 JFK STREET
CAMBRIDGE, MA 02138

GROWTHLAB.HKS.HARVARD.EDU

Acknowledgments
The authors would like to thank Ostap Stefak for excellent research assistance and Sebastian Bustos and Muhammed Yildirim for valuable comments. We are further grateful for the comments received at the Global Conference on Economic Geography (Dublin), Geography of Innovation (Milan, Manchester), European Regional Science Association (Alicante), the workshop on Economic Complexity, Geography and Innovation (Cambridge, MA), the Economic Fitness and Complexity Spring School (Rome), and at seminars at the Harvard Growth Lab (Cambridge, MA) and the IPP-CSIC (Madrid).

This paper may be referenced as follows: Schetter, U., Diodato, D., Protzer, E., Neffke, F. and Hausmann, R. (2024). “From Products to Capabilities: Constructing a Genotypic Product Space.” Growth Lab Working Paper, Harvard University Kennedy School of Government.

From Products to Capabilities: Constructing a Genotypic Product Space

Ulrich Schetter*, Dario Diodato*, Eric Protzer,
Frank Neffke, Ricardo Hausmann†

June 2024

Abstract Economic development is a path-dependent process in which countries accumulate capabilities that allow them to move into more complex products and industries. Inspired by a theory of capabilities that explains which countries produce which products, these diversification dynamics have been studied in great detail in the literature on economic complexity analysis. However, so far, these capabilities have remained latent and inference is drawn from product spaces that reflect economic outcomes: which products are often exported in tandem. Borrowing a metaphor from biology, such analysis remains phenotypic in nature. In this paper we develop a methodology that allows economic complexity analysis to use capabilities directly. To do so, we interpret the capability requirements of industries as a genetic code that shows how capabilities map onto products. We apply this framework to construct a genotypic product space and to infer countries' capability bases. These constructs can be used to determine which capabilities a country would still need to acquire if it were to diversify into a given industry. We show that this information is not just valuable in predicting future diversification paths and to advance our understanding of economic development, but also to design more concrete policy interventions that go beyond targeting products by identifying the underlying capability requirements.

Keywords product space · economic complexity · economic convergence · export diversification · industrial policy · poverty trap · structural change

JEL Classification O11 · O14

*Ulrich Schetter and Dario Diodato contributed equally to this paper. The order among these two authors has been drawn at random.

†Ulrich Schetter: University of Pavia (ulrich.schetter@unipv.it); Dario Diodato: Joint Research Centre, European Commission (dario.diodato@ec.europa.eu); Eric Protzer: Growth Lab, Harvard University; Frank Neffke: Complexity Science Hub Vienna; Ricardo Hausmann: Growth Lab, Harvard University and Santa Fe Institute. The authors would like to thank Ostap Stefak for excellent research assistance and Sebastian Bustos and Muhammed Yildirim for valuable comments. We are further grateful for the comments received at the Global Conference on Economic Geography (Dublin), Geography of Innovation (Milan, Manchester), European Regional Science Association (Alicante), the workshop on Economic Complexity, Geography and Innovation (Cambridge, MA), the Economic Fitness and Complexity Spring School (Rome), and at seminars at the Harvard Growth Lab (Cambridge, MA) and the IPP-CSIC (Madrid).

1 Introduction

Economic development is often cast as a process of structural transformation in which countries diversify by entering new economic activities. A literature that goes back to at least Kim (1980) and Abramovitz (1986) has argued that, to transform the structure of their economies, countries need specific "capabilities". Ever since, scholars have endeavored to operationalize the notion of capabilities empirically (Archibugi and Coco, 2005; Fagerberg and Srholec, 2008). This undertaking has proved fraught with difficulties: it requires an exhaustive list of capabilities, empirical strategies to measure them and weights that determine each capability's importance. Recently, an alternative approach has emerged in a field that we will refer to as *economic complexity analysis* (ECA). This field considers capabilities to be pivotal determinants of the industrial structure of economies: countries produce the goods and services for which they have all prerequisite capabilities. Based on this reasoning, Hidalgo et al. (2007) propose that one can assess which products require similar capabilities by observing which products are often produced by the same countries. This approach has been successfully applied to predict diversification trajectories of countries, regions and cities, not only in terms of their economic output, but also in the technological and scientific areas they are able to enter (Hidalgo et al., 2018).

While this notion of similarity has an implied technological root – the degree of similarity among products must in some way be related to their capabilities – the measurement of the product space abstracts from how products are made and focuses, instead, on what is more readily observable: the final output of countries. To borrow a metaphor from biology, we argue that this approach is *phenotypic* in nature: it connects products not by similarities in their "DNA", i.e., the capabilities they require, but by the way this DNA is expressed in the mix of products that countries export.[1] In spite of its predictive utility, this phenotypic approach makes it hard to ask a number of important questions about economic development and development policy, such as: Which capabilities does a country have? Which products are feasible with this set of capabilities? Which capabilities does the country need to acquire to enter a specific new economic activity? And: are some capabilities more easily acquired than others? To overcome these deficiencies, we build on previous work by Hausmann and Hidalgo (2011), and Diodato et al. (2022) to propose a

[1] Although we borrow this terminology from evolutionary theory developed in biology, we do not take a strong position in the debate on generalized Darwinism (e.g., Aldrich et al., 2008). In fact, we will develop an empirical methodology that we hope will prove useful to test and develop a broad range of theoretical frameworks.

tractable approach to constructing a *genotypic* product space. Doing so, we will not only map the capability requirements of products, but also infer the capability endowments of countries, only using widely available data. This, in turn, offers a new view on a country's capability base, as well as the opportunities and challenges the country faces on its future development path.

The central idea we leverage is that one can interpret the capability requirements of industries as something like a genetic code, a mapping from capabilities to products. This allows augmenting previous approaches to the product space in three ways: first, we can calculate the distance between two industries by counting the number of capabilities that are required in one, but not in the other. Second, by focusing on non-tradeable inputs and assuming strong limits to substitution among them, we can infer which such inputs are available in a given economy. In particular, under such assumptions, a country can only export a product if it possesses all non-tradeable inputs (or, *capabilities*) that the product requires. Third, once input requirements of products and input endowments of countries are known, we can directly compute the *genotypic* proximity between any product and any country. This allows determining which capabilities need to be acquired to render specific diversification paths feasible.

We apply this methodology focusing on capabilities embedded in the workforce and showcase how these capabilities help predict and understand the diversification of countries' export portfolios. We focus on capabilities connected to human capital, because the plausibility of our framework hinges on the assumption that inputs are non-tradeable and non-substitutable. These conditions are likely to be approximately fulfilled for human capital because, on the one hand, in many jobs, human capital is highly specific (for instance, there is no reasonable rate of substitution between car engineers and accountants) and, on the other hand, workers' mobility is strongly constrained by geographic distance and country borders. However, the framework itself allows the use of any type of input that can be considered a capability in the aforementioned sense.

We proceed as follows. First, we construct a matrix that describes the occupational requirements for each industry in the economy, using the US Bureau of Labor Statistics' Occupational Employment Statistics. We use the resulting *capability requirements matrix* to construct a *genotypic product space* and show how this space has certain advantages over its phenotypic counterpart. Next, we combine the capability requirements matrix with data on countries' exports to infer the capability endowments of countries. We test the validity of our genotypic framework in an analysis of how countries diversify their

export baskets. Finally, we discuss implications.

Doing so reveals numerous conceptual advantages linked to the genotypic approach's clear interpretation of what it means that two products are related. These advantages express themselves in more informative descriptions of countries' capability bases, as well as of the developmental bottlenecks they imply. Empirically, we show that our genotypic proximity has comparable predictive performance to standard phenotypic approaches, while providing a more linear mapping between proximity and diversification probabilities. Furthermore, we show how genotypic proximities can be augmented by incorporating country-product specific information on the complexity of missing capabilities, which further improves predictions. Finally, we show how the genotypic approach can be used in policy-making and sketch an agenda for future research in a genotypic approach to ECA.

2 Literature

Our work is inspired by and complements a vast body of research on structural transformation and catching up in economic development (Abramovitz, 1986; Hirschman, 1958; Lall, 1992; Kim, 1980; Fagerberg et al., 2010). A central question in this literature is why productivity differs so widely across economies and a core explanation is that countries differ in their *technology*. Since Abramovitz' (1956) assertion that the Solow residual is nothing but a "measure of our ignorance", a group of scholars has tried to capture a country's state of technology by studying social (Abramovitz, 1986) and technological (Kim, 1980) *capabilities*. This set in motion a broad effort in evolutionary economics (Nelson and Winter, 1982) to identify and measure these capabilities. However, such an endeavor faces several challenges (see, for instance, Fagerberg et al., 2010). First, modern economies typically rely on a wide variety of capabilities.[2] Second, capabilities are often inherently difficult to observe, because they have tacit (Polanyi, 1962) components. Third, even if we had a more or less exhaustive list of capabilities and ways to identify them, we would still need to determine how to avoid double-counting closely related capabilities and how to weigh capabilities according to their importance.

An alternative approach to understanding economic development is formulated in New Structuralist Economics (NSE, Lin, 2011). Building on "old" structuralist economics

[2]For instance, Lall (1992) considers three broad classes of capabilities: physical investment capabilities (related to, for instance, the financial sector), human capital (related to health, schooling and training) and technological capabilities (related to research, innovation and commercialization).

(e.g., Hirschman, 1958; Prebisch, 1962), NSE focuses on structural transformation. It argues that productivity and future development prospects are intimately linked to the type of activities that economies engage in (Hausmann et al., 2007), with self-sufficient agriculture at the bottom rungs of the developmental ladder and industrial activities in machinery and electronics, as well as advanced business services, at the top, akin to economic models with a 'ladder' of development (Krugman, 1985; Lucas, 1993; Hausmann and Rodrik, 2003; Costinot, 2009; Sutton and Trefler, 2016; Schetter, 2020; Atkin et al., 2021). Countries cannot freely choose their activities. Instead, they specialize according to their comparative advantage, which depends on their factor endowments, and so does their potential for structural transformation.[3] This includes the traditional factors of land, capital and labor. Other important factors are influenced by government actions: the economy's so-called hard and soft infrastructure (Lin, 2011), where the former refers to physical infrastructure (roads, ports, electricity grids), whereas the latter includes less tangible infrastructure, such as institutions, universities, or financial regulation. These factor endowments bear a striking resemblance to the capabilities identified in evolutionary economics. However, whereas the capability literature mostly focused on the link between countries' composite capabilities and aggregate growth (for instance, as expressed in their GDP per capita), NSE pays special attention to how specific factor endowments facilitate the development of some sectors but not others.[4]

NSE studies structural transformation in broad categories, both in terms of factor endowments and the sectors they support – agriculture, heavy industry, high tech industries, etc.. In this paper, we instead build on a closely related field, economic complexity analysis (ECA). Like NSE, ECA starts from the assumption that different activities require different capabilities. However, these capabilities tend to be more fine-grained. Furthermore, ECA assumes strong complementarities among capabilities, with little room for substitution between them. As a consequence, economies can only produce the products for which they possess all required capabilities. Just like the production factors of NSE, capabilities can be physical, like specific pieces of equipment or infrastructure, or intangible, as in the case of institutions or technological expertise. However, not all capabilities matter equally in determining which products can be produced where, and ECA therefore

[3]The NSE paradigm is thus related to a voluminous literature on structural change (e.g. Kuznets 1957; Kongsamut et al. 2001; Foellmi and Zweimüller 2008; Buera et al. 2022) and in particular papers analyzing structural change in open economies (e.g. Matsuyama 1992; Uy et al. 2013; Matsuyama 2019), where NSE puts a strong emphasis on the 'capabilities' that drive structural transformation (see below).

[4]Another difference is that evolutionary economists, emphasizing innovation, often concentrate on technological as opposed to production capabilities, whereas the new structuralist approach of Lin does not give preference to one over the other.

focuses on capabilities that meet a number of criteria: they should be non-ubiquitous, hard to access from outside the economy (i.e., they should be non-tradable), but relatively easy to access by different firms within the economy.[5]

Even with these restrictions, ECA shares the core methodological challenges of the earlier capability-based approaches, namely capabilities' high multiplicity, limited observability and unknown weights. Therefore, a crucial methodological innovation of ECA is that it bypasses enumerating capabilities and identifying capability requirements of products and capability endowments of economies. To do so, ECA developed abstract networks based on co-occurences that express similarities in capability requirements in so-called product (Hidalgo et al., 2007), industry[6] (Neffke et al., 2011), technology (Kogler et al., 2015) or multilayer spaces (Pugliese et al., 2019), and similarities in capability endowments in country (Bahar et al., 2014) or city spaces. Such spaces are highly predictive of diversification patterns—see also Hausmann and Klinger (2006); Boschma and Frenken (2006); Frenken and Boschma (2007); Hidalgo et al. (2018).[7] We start from a conceptual framework that links products to underlying capabilities (Hidalgo and Hausmann, 2009; Hausmann and Hidalgo, 2011; O'Clery et al., 2021). We then add to the literature by exploiting the underlying capability structure directly to learn about the product space and related diversification. Within the wider literature in economic geography, our work thus also relates to efforts to understand co-agglomeration patterns of industries (Ellison et al., 2010; Diodato et al., 2018; Steijn et al., 2022) or co-exporting patterns of products (Bahar et al., 2019). Such co-agglomeration and co-exporting patterns are nothing else than industry or product spaces in the parlance of economic complexity analysis and our genotypic approach may therefore also offer new ways to shed light on the drivers of agglomeration externalities. Finally, the various implications for policy of our work contribute to an expanding set of papers that explores how economic complexity analysis can be used as a policy framework (Hidalgo, 2023; Balland et al., 2018; Boschma et al., 2021; Li and Neffke, 2023).

[5]See Neffke et al. (2018) for a more complete exposition as well as similarities to the notion of sustained competitive advantage (Barney, 1991) in management science.

[6]The idea of an industry space can be traced to the management literature (Teece et al., 1994), where scholars were confronted with the same obstacles to identify and measure the resource bases of firms.

[7]Furthermore, to get a sense of the extent of an economy's capability base, it developed metrics of economic complexity, i.e., estimates of the completeness of an economy's capability endowments (Hidalgo and Hausmann, 2009; Tacchella et al., 2012).

3 Capabilities-based view on production

At the core of ECA is the idea that modern production relies on many distinct capabilities, and that products require overlapping but distinct subsets of these capabilities. These capabilities are necessary inputs, i.e., making a given product entails acquiring the entire set of capabilities that this product requires. This model of production can be succinctly represented in matrix form (Hausmann and Hidalgo, 2011). To that end, let $\boldsymbol{C}$ be a $N_c \times N_a$ dimensional binary capability-endowment matrix, where N_c is the number of countries and N_a the number of capabilities. An entry of this matrix, C_{ca}, equals one if country c has capability a and zero otherwise. Similarly, let $\boldsymbol{P}$ be a $N_p \times N_a$ dimensional capability-requirements matrix—where N_p is the number of products—whose elements, P_{pa}, indicate whether capability a is required to produce product p.

Together, the capability-endowments matrix $\boldsymbol{C}$ and the capability-requirements matrix $\boldsymbol{P}$ tell us which countries can make which products. In particular, country c can only make product p if $\sum_a (1 - C_{ca})P_{pa} = 0$.[8] We can collect this information in a binary $N_c \times N_p$ matrix, $\boldsymbol{M}$, that describes which countries can make which products:

$$M_{cp} = \mathbb{1}\left[\sum_a (1 - C_{ca})P_{pa} = 0\right] \tag{1}$$

where $\mathbb{1}[]$ is the indicator function that evaluates to 1 if the term in brackets is true and zero otherwise. If countries make all the products they possibly can—a common (implicit) assumption in the related literature—$\boldsymbol{M}$ will correspond to a matrix that represents countries' specialization in international trade.[9]

$\boldsymbol{P}$ is binary and elements P_{pa} simply indicate whether or not the capability a is needed to produce the product p. In what follows, we will also consider a variant of the capability-requirements matrix, $\tilde{\boldsymbol{P}}$. This matrix has elements between 0 and 1 that describe how *intensively* product p makes use of capability a. Specifically, we can think of elements $\tilde{P}_{pa}$ as representing cost shares such that $\sum_a \tilde{P}_{pa} = 1$ for any p. Note that $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ are related: whenever $\tilde{P}_{pa} > 0$, $P_{pa} = 1$ and whenever $\tilde{P}_{pa} = 0$, $P_{pa} = 0$.

[8] The term $(1 - C_{ca})$ evaluates to one whenever a country c does not have capability a, such that the summation only equals zero if a country misses none of the capabilities that product p requires.

[9] A large literature in international trade suggests that countries should specialize according to their comparative advantage. Nevertheless, standard multi-industry (or -product) gravity models do not predict zeros at the exporter-industry level, i.e., there is no specialization at the extensive industry margin (see, e.g. Costinot et al. 2012). More importantly, this simplification is in line with the fact that even relatively small rich countries like Portugal, Czech Republic or Denmark export more than 95% of the ∼1200 products at the 4-digit HS-level. It is also in line with the observation that exports tend to be 'nested' (Hausmann et al., 2011; Bustos et al., 2012; Tacchella et al., 2012; Schetter, 2020; Gersbach et al., 2023).

3.1 Phenotypic approach

The previous discussions suggest that much can be learned about the underlying capability structure without knowing the underlying matrices $\boldsymbol{C}$ and $\boldsymbol{P}$ and by instead focusing on the economic outcomes matrix $\boldsymbol{M}$. This is the approach taken by the bulk of the literature on economic complexity analysis. The basic idea is simple: if capabilities represent necessary, non-tradeable inputs (broadly defined), then the fact that a country makes a product means that the country must have all required capabilities or, equivalently, that the product requires only capabilities that are available in the country. Consequently, products that require similar capabilities are likely to be exported by the same countries. This reading implies that co-exporting patterns reveal which products have similar capability requirements, an idea that motivated the construction of the product space (Hidalgo et al., 2007)—a network representation that connects products if they are often co-exported.

The elements of $\boldsymbol{M}$ describe which countries export which products. In practice, they are often determined by calculating a country's revealed comparative advantage (RCA, or Balassa index Balassa (1965)) in a product:

$$RCA_{cp} = \left(x_{cp} / \sum_{p} x_{cp} \right) / \left(\sum_{c} x_{cp} / \sum_{c} \sum_{p} x_{cp} \right), \tag{2}$$

where x_{cp} represents the value of the exports of country c in product p. The RCA compares the share of product p in country c's exports to p's share in global exports. Values over one indicate that the country is specialized in product p and hereafter we will say that country c (significantly) exports product p if $RCA_{cp} > 1$ and set $M_{cp} = 1$ in such case and 0 otherwise.

The product space is constructed based on measures of *proximity* between pairs of products which Hidalgo et al. (2007) define as:

$$\Phi_{pp'} = \frac{\sum_c M_{cp} M_{cp'}}{\max(\sum_c M_{cp}, \sum_c M_{cp'})}. \tag{3}$$

There are many variations in how to calculate this proximity (Li and Neffke, 2023), the details of which do not matter for our purposes. The key point is that these measures all rely only on matrix $\boldsymbol{M}$, i.e., on observed outputs, not on information about actual capabilities. For this reason, we refer to this class of measures as *phenotypic* product spaces.

The product space can be used to predict the evolution of the $\boldsymbol{M}$ matrix, that is, the diversification of countries into new products. Intuitively, if building up new capabilities is costly, then it should be easier for a country to move into nearby products, that is, products that require few new capabilities. Such products should be close to the country's current activities where proximity is defined in terms of the topology of the product space.

Given that the product space is based on similarities between products, we need to translate these pairwise similarities into a similarity between a country—i.e., a basket of products—and a potential target product. Typically, this is achieved by assessing how active the country is in products closely related to this target product, or, by determining the "*density*" of an economy around each product.[10] For instance, Hidalgo et al. (2007) calculate the density of country c around product p, ω_{cp}, as:

$$\omega_{cp} = \frac{\sum_{p'} M_{cp'} \Phi_{pp'}}{\sum_{p'} \Phi_{pp'}}. \tag{4}$$

Density measures have proved remarkably predictive of diversification processes well beyond the export portfolios of countries. Related diversification is so prevalent that it has been coined the *principle of relatedness* (Hidalgo et al., 2018). This empirical success and the minimal data requirements explain why keeping the analysis at the phenotypic level has had such an appeal, specially given that capability endowments and requirements are difficult to observe and could exhibit complex and heterogeneous structures.

However, the phenotypic approach also has several shortcomings. First, there are several *ad hoc* choices in the design of product spaces and density metrics (Li and Neffke, 2023). Second, phenotypic proximity measures are symmetric while the underlying capability structure often implies a directionality. Intuitively, it should be easier to move from motorcycles to bicycles than vice-versa. Third, density measures may double-count capabilities if products close to the focal product are also closely related to one another. Fourth, density measures are not informative about *which* capabilities a country lacks that prevent it from entering the economic activity. In other words, phenotypic measures have much to say about which products to diversify into, but much less about how to get there. This limits their use in devising concrete development policies. Fifth, and related to this, density measures do not help distinguish between capabilities that may differ in importance or how hard it is to acquire them.

[10] Again, density measures can be constructed in various ways (see Li and Neffke, 2023, for an overview), but all rely on product-product similarities to arrive at an estimate of how close a single product is to a set of products.

3.2 Genotypic approach

To remedy the shortcomings of the phenotypic approach in ECA, we build on Diodato et al. (2022) and develop a genotypic alternative that results in genotypic proximity and density measures. In this alternative approach we aim to develop a window directly on a country's capability base. Doing so requires that we can observe the capability requirements of products, i.e., matrix $\tilde{\boldsymbol{P}}$ or $\boldsymbol{P}$. We will discuss the measurement of these matrices in Section 4. Here, we focus instead on the conceptual framework, and to that end we simply assume for now that we are equipped with matrix $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}$, respectively.

3.2.1 Genotypic proximity

Given matrices $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}$, respectively, we can measure the technological proximity between two products by directly comparing their capability requirements. In particular, consider the following measure of proximity between products p and p', $\Gamma_{pp'}$:

$$\Gamma_{pp'} = \frac{\sum_a (P_{pa} P_{p'a})}{\sum_a P_{pa}}. \tag{5}$$

The numerator of eq. (5) counts the number of capabilities that are required in both products, p and p', whereas the denominator counts the total number of capabilities required in product p. Consequently, $\Gamma_{pp'}$ indicates which share of the capabilities that are needed to produce p is also used to produce p'. That is, $\Gamma_{pp'}$ focuses only on the extensive capability margin by treating all capabilities symmetrically. Alternatively, we can weigh capabilities by how intensively they are used in the production of p, i.e., by their cost shares (for instance, when we know that the cost share of a_1 in product p is twice as large compared to a_2):

$$\tilde{\Gamma}_{pp'} = \sum_a (\tilde{P}_{pa} P_{p'a}). \tag{6}$$

$\tilde{\Gamma}_{pp'}$ now indicates the total cost-share in p of capabilities that are also required by product p'. Note that we can convert these proximity measures into measures of pairwise distance as $1 - \Gamma_{pp'}$ and $1 - \tilde{\Gamma}_{pp'}$, respectively.

Unlike the phenotypic product space, the genotypic product space is not inferred but directly constructed from information about capability requirements of products. To see the merit of these measures, suppose that acquiring a capability entails a fixed cost f per capability. Then, $1 - \Gamma_{pp'}$ indicates the cost of starting to produce p when a country already produces p' and, hence, has all the capabilities needed for p'. Instead, if the cost

of acquiring a new capability is proportional to how intensively it is used in the target product—e.g. because of an initial learning phase (Diodato et al., 2022)—, this cost is captured by $1 - \Gamma_{pp'}$. Note that these measures are directed, i.e., in general, $\Gamma_{pp'} \neq \Gamma_{p'p}$. This direction captures the fact that it is easier to move from a complex product to a closely related, but less complex product than vice versa. For instance, if product p_1 uses capabilities a_1 and a_2, while product p_2 only uses a_1 moving from p_1 to p_2 should be easier than moving from p_2 to p_1.

3.2.2 Inferring matrix $\boldsymbol{C}$

To use the genotypic approach for analyzing diversification patterns, we need to further know the capability endowments of countries. Given that not all countries have equally good data and that existing data are rarely harmonized across countries, this is a complex undertaking. However, as shown in Diodato et al. (2022), we can leverage equation (1) to infer the capabilities of countries in matrix $\boldsymbol{C}$ from matrices $\boldsymbol{P}$ and $\boldsymbol{M}$. This equation states that a country can make a product only if it has all required capabilities. This, in turn, allows inferring the country's capability endowments from the products it makes. For instance, if producing engines requires mechanical engineering know-how, the fact that a country makes engines implies that this know-how is part of the capability endowments of that country. More generally, we can infer matrix $\boldsymbol{C}$ from the production matrix $\boldsymbol{M}$ and the capability requirements matrix $\boldsymbol{P}$ as follows

$$C_{ca} = \mathbb{1}\left[\sum_p M_{cp} P_{pa} > 0\right], \tag{7}$$

where the term $\sum_p M_{cp} P_{pa}$ counts how many products produced by country c use capability a. If this sum is strictly positive, country c must have capability a.

3.2.3 Genotypic density

The key advantage of inferring $\boldsymbol{C}$ is that it allows computing the proximity of a country to a product in a way that is consistent with the underlying framework outlined at the beginning of Section 3:

$$\mu_{cp} = \frac{\sum_a C_{ca} P_{pa}}{\sum_a P_{pa}}. \tag{8}$$

Eq. (8) measures which share of capabilities that product p requires country c already owns. Alternatively, we can derive a density metric that factors in how intensively the different capabilities are used in product p, analogously to eq. (6):

$$\tilde{\mu}_{cp} = \sum_a C_{ca} \tilde{P}_{pa}. \tag{9}$$

These measures can again be translated into distances by subtracting them from 1. $1-\mu_{cp}$, for example, shows which share of capabilities country c would still have to acquire in order to start producing p. If the cost of acquiring capabilities scales with the intensity of their use—for instance, when the quality or productivity of a capability is smaller in an initial learning phase—, then $1-\tilde{\mu}_{cp}$ is more appropriate. This distance reflects how intensively a product relies on capabilities that country c still lacks. Either distance metric determines the ease with which country c can enter product p (Diodato et al., 2022).

The genotypic density metrics account for the full proximity relations between all pairs of products while the phenotypic density of eq. (4) does not as it sums all pairwise proximities of product p to any other product p'. Consider, for example, the extreme case where two products p' and p'' have the exact same capability requirements. Then, if a country makes product p', the fact that it also makes p'' does not add any information about its underlying capability endowments. Hence, it does not help determine how close the country is to a candidate product p. Nevertheless, standard phenotypic density metrics will suggest that the country is closer to product p when it exports both products, p' and p'', than when it exports only one of them. By contrast, the genotypic density accounts for such duplicities.[11]

4 Data

To put the genotypic approach to the test, we study its performance in the canonical application of ECA to international trade. To do so, we rely on the U.N. Comtrade's data on exported commodities between 1992 and 2016 (*trade data*).[12] We add to this information on occupational profiles of industries from the Occupational Employment Statistics for year 2002 compiled by the US Bureau of Labor Statistics (*BLS data*).[13]

We start by creating the matrix $\boldsymbol{M}$. To do so, we follow Hidalgo et al. (2007) and create a binary matrix that describes which products are significantly present in the export basket of which countries, based on the RCAs of eq. (2). That is, we assign to each element M_{cp} of matrix $\boldsymbol{M}$ a value of 1 when RCA_{cp} is greater than one and a value of 0 otherwise.

Next, we need an estimate of matrix $\boldsymbol{P}$ (or of $\tilde{\boldsymbol{P}}$). That is, we need a description of which

[11] Formally, observe from eq. (7) that for the inferred matrix $\boldsymbol{C}$ it makes no difference if a capability is used in only one or several of a country's products.

[12] We use a version of the dataset that has been processed for the Atlas of Economic Complexity (https://atlas.cid.harvard.edu/about-data).

[13] Occupational Employment Statistics (OES) (now Occupational Employment and Wage Statistics).

capabilities are used to produce which products. We will assume that, to a first approximation, these capability requirements data are constant over time and across countries. That is, we will assume that products are made in the same way across the globe and across different time periods. This assumption is rather restrictive. However, similar assumptions are typically made about phenotypical product spaces. Moreover, it is easy to allow for heterogeneity in capability requirements. For instance, one could use a different matrix for developing countries or allow capability requirements to change over time. To keep the exposition simple, we will here focus on the simplest case where $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ are universal, leaving other scenarios for future research.

Capabilities may come in many different forms, such as specialized skills, technological know-how, infrastructure, or institutions. Here, we limit ourselves to a particularly important class of capabilities: the capabilities that are embedded in human capital, the skills and know-how of the workforce needed to work in different occupations. On the one hand, access to skilled workers fulfills important aspects of capabilities: human expertise is geographically sticky and different types of expertise are often poor substitutes for one another. On the other hand, human capital data is often readily available from labor market surveys or censuses.

Specifically, we construct for each industry occupational employment vectors, based on information from BLS data. These vectors tell us for each industry in the US, which share of its wage bill goes to workers in any given occupation. To merge these data with our trade data, we use a concordance developed by Pierce and Schott (2009) that links 6-digit HS commodity codes to 4-digit NAICS industry codes.[14] Finally, we drop countries with fewer than 2 Million inhabitants.[15] This yields a dataset with 140 countries, 88 industries and 444 occupations.

Next, we assign a value of 1 to element P_{pa} of matrix $\boldsymbol{P}$ if occupation a is used by industry p. Similarly, for matrix $\tilde{\boldsymbol{P}}$, we set element $\tilde{P}_{pa}$ equal to the wage-bill share of occupation a in product p.[16] We then use these matrices to infer capability endowments of countries, as described in eq. (7).

Matrix $\boldsymbol{M}$ has been the object of study of the phenotypic approach to ECA. First-order quantities that can be calculated from this matrix are the ubiquity of products and the diversity of country's export baskets. The first rows of Table 1 provide some information

[14] See Diodato et al. (2022) for further details.

[15] We use population data from Feenstra et al. (2015).

[16] We also considered employment shares, yielding qualitatively the same results as the ones presented here. Details are available upon request.

Table 1: Descriptive statistics for M_{cp}, P_{pa}, and C_{ca}

Var.	N	Mean	St.Dev	Min/Max	
				M_{cp}	
M_c	140	20.73	13.03	min: 1 -	Angola
				max: 56 -	Poland
M_p	88	32.98	14.22	min: 11 -	Commercial and Service Industry Machinery
				max: 67 -	Other Food
				P_{pa}	
P_p	88	126.19	35.73	min: 39 -	Leather and Hide Tanning
				max: 204 -	Navigational and Electromedical Instruments
P_a	444	25.01	28.74	min: 1 -	Air Traffic Controllers
				max: 88 -	Bookkeeping, Accounting, and Auditing Clerks
				C_{ca}	
C_c	140	301.83	54.39	min: 138 -	Angola
				max: 397 -	Netherlands
C_a	444	95.17	43.40	min: 12 -	Air Traffic Controllers
				max: 140 -	Accountants and Auditors

Notes: Descriptive statistics are reported for the following variables: M_c (country diversity = $\sum_p M_{cp}$); M_p (industry ubiquity = $\sum_c M_{cp}$); P_p (industry span = $\sum_a P_{pa}$); P_a (capability generality = $\sum_p P_{pa}$); C_c (country completeness = $\sum_a C_{ca}$); C_a (capability dispersion = $\sum_c C_{ca}$).

based on these quantities. The most *diversified* country is Poland, exporting 56 out of 88 industries. Angola, on the other hand, is only active in one industry (Oil and Gas Extraction). When it comes to industries, the least *ubiquitous* industry is the *Commercial and Service Industry Machinery*, with only 11 countries that export it, while the most ubiquitous industry is *Other Food*.

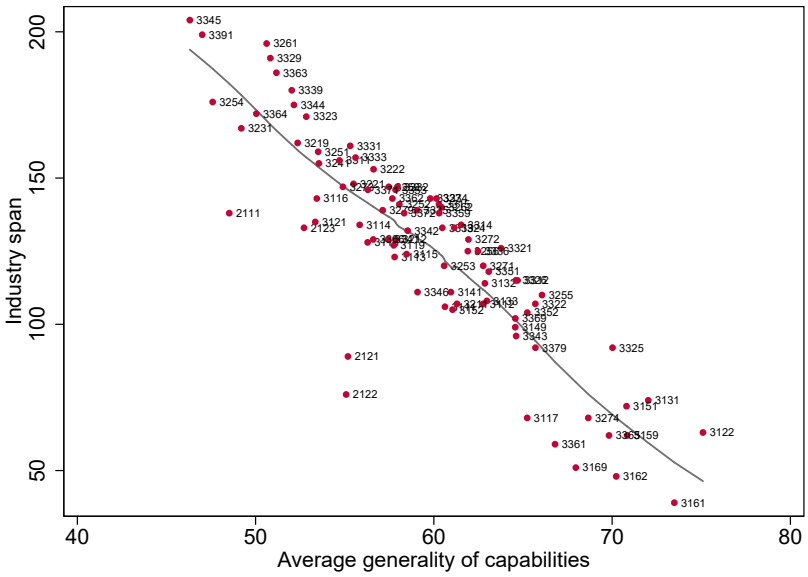
The genotypic approach offers two additional matrices that can be studied, $\boldsymbol{P}$ and $\boldsymbol{C}$. These matrices describe capability requirements of industries and capability endowments of countries, respectively. Based on these matrices we can ask, for instance, which industries have the shortest capability *span* (*Leather and Hide Tanning*) and which the longest (*Navigational and Electromedical Instruments*). Similarly, we see a large variation in the *completeness* of countries' capability endowments, ranging from Angola, with just 138 occupations to the Netherlands, where industries can access over 400 occupations.

We can also ask how capabilities are distributed across industries and countries. Capabili-

ties run from highest *generality*, such as the skills of *Accounting Clerks*, who are employed in each of the 88 industries, to the highly specific skills of *Bookbinders*, who work in only a single industry. When looking into the *dispersion* of capabilities across countries instead, we see that the general skills of *Accountants* are found in every country, whereas the expertise of *Air Traffic Controllers* is much more concentrated.[17] In Appendix A.1, we provide further rankings along each of these dimensions.[18]

The genotypic approach also allows us to study how these quantities relate to one another. As an example, Fig. 1 shows that the capability span of an industry is strongly and negatively correlated with the average generality of capabilities. That is, industries that require many capabilities typically also rely on highly specific capabilities. This suggests that the development process entails not only the accumulation of more but also of ever more specialized capabilities.

Figure 1: Industry span and average generality of capabilities in the industry



Notes: The y-axis depicts the number of occupations required by an industry, the industry's *capability span*: $P_p = \sum_a P_{pa}$. The x-axis shows the industry's average *generality*, i.e., the average number of industries in which the capability is required.

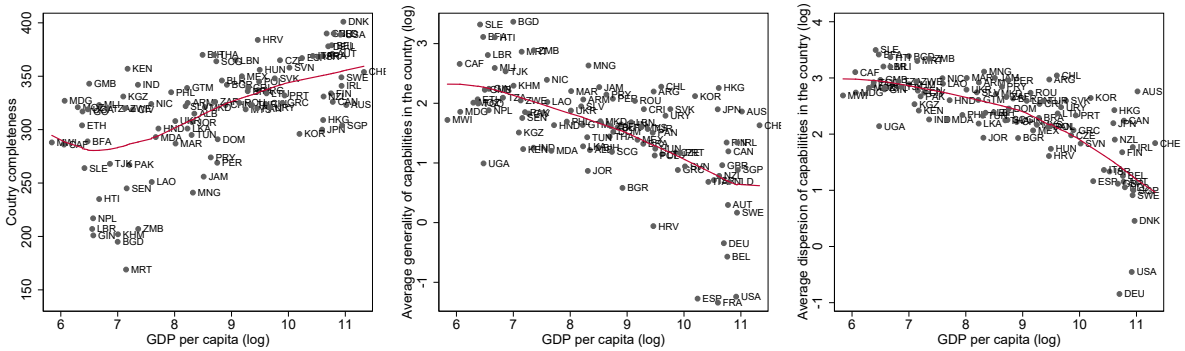
Similarly, Figure 2 shows how capability bases vary across countries with different levels of development. Panel a shows how the completeness of a country's capability base correlates with the country's GDP per capita. Richer countries tend to have more capabilities,

[17]Note that services are not included in our sample. Consequently, *Air Traffic Controllers* are only required in the *Aerospace Products and Parts* industry that is active in only 12 countries.

[18]The minimums and maximums in Table 1 are often a tie with other countries, products or occupations. For instance, both *Accounting Clerks* and *Accountants* can be found in all 140 countries.

Figure 2: Capabilities in the development process

(a) Completeness (b) Generality (c) Dispersion



Notes: The x-axis represents GDP per capita (in logs), while the y-axis shows: (a) country completeness (the revealed number of occupations in the country); (b) the log of the average generality of capabilities in the country; (c) the logs of average dispersion of capabilities in the country.

consistent with the fact that they also tend to have more diversified export baskets.[19] Furthermore, panel b shows that, on average, richer countries have more specific capabilities and panel c shows that these capabilities are more concentrated in a small set of countries. Together, these plots suggest that development entails not only accumulating more but also more specialized capabilities that go into fewer products and are present in fewer countries.

5 Diversification dynamics

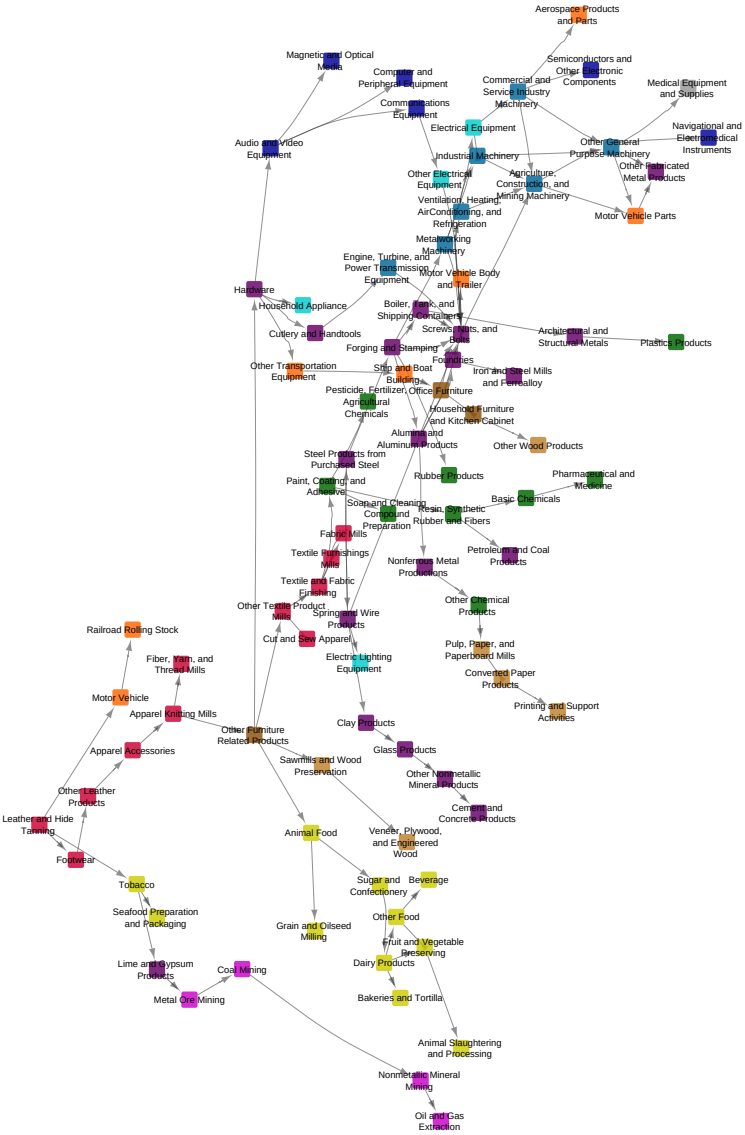
5.1 The occupation-based genotypic product space

Figure 3 visualizes the genotypic product space $\tilde{\mathbf{\Gamma}}$ as a network. In this network, each node represents an industry, labelled with an abbreviated name and colored according to the 3-digit Naics sector to which they belong.

The positioning of nodes deviates from how phenotypic industry spaces are commonly displayed, which typically use force-directed algorithms. Instead, Figure 3 arranges nodes along the vertical axis to enhance visual clarity by limiting edge crossings, whereas the

[19] Given that we infer capabilities from a country's pattern of specialization and capability requirements in the US, these two observations are closely related. Although this approach is not that different from the common assumption in ECA of a universal product space that does not differ across countries. To assess how well the assumption of a universal capability matrix is we would need comparable data on input requirements by product. Such an analysis is however beyond the scope of the current paper.

Figure 3: Genotypic product space, $\tilde{\Gamma}_{pp'}$



Notes: Each node represents a 4-digit NAICS industry, positioned on the horizontal axis according to the number of occupations it requires. Color coding: *red*=textiles; *yellow*=food processing; *light purple*=extraction; *dark purple*=manufacturing of mineral and non-mineral products; *green*=chemicals; *light brown*=paper and wood products; *dark brown*=furniture; *orange*=transportation; *light blue*=electrical equipment; *teal*=machinery; *dark blue*=electronics; *grey*=others.

horizontal axis sorts industries in ascending order of their capability span. That is, industries that require the smallest number of capabilities are situated to the left (*Leather and Hide Tanning*, *Footwear*, and *Other Leather Products*) and highly complex industries that require many capabilities are situated to the right (*Navigation and Electromedical Instruments*, *Medical Equipment*, and *Plastic Products*).

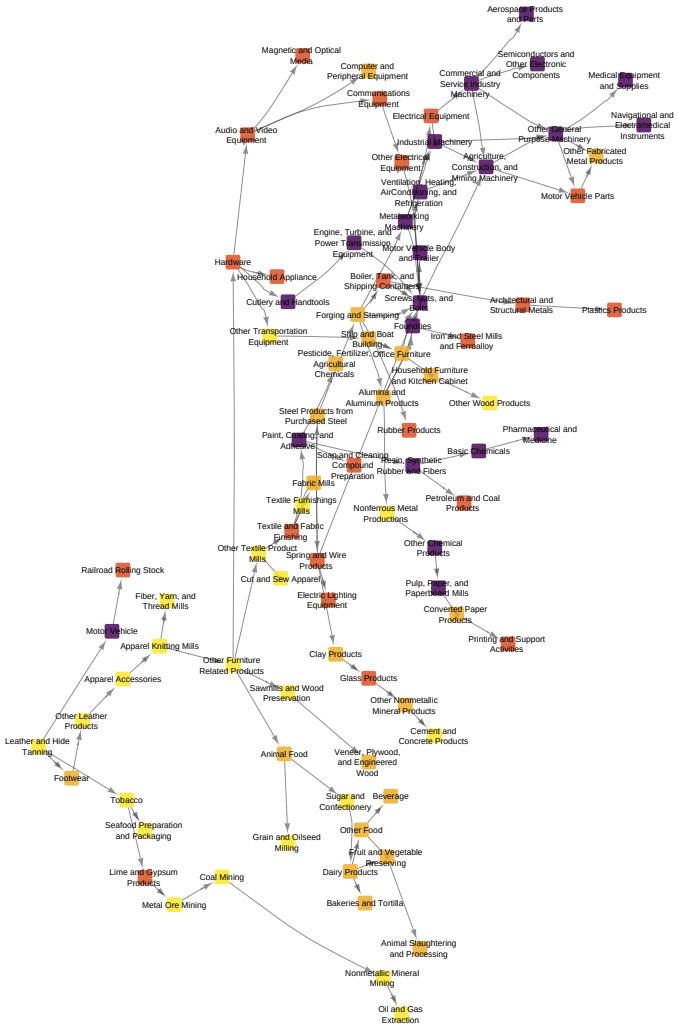
Links between nodes represent the proximity between industries as measured by $\tilde{\Gamma}_{pp'}$. Unlike the phenotypic product space, edges are directed and we here only show arrows that point right towards more complex industries. To avoid cluttering the visualization, we only draw edges where $\tilde{\Gamma}_{pp'} > .98$. Finally, to avoid isolated nodes, where needed, we add the closest incoming connection for each industry.

Interestingly, Fig. 3 suggests that there is a single high-level development path that takes countries from the most basic industry, *Leather and Hide Tanning*, to more complex industries. This path runs via textiles (red) into chemicals (green) and machinery (teal), to end in the most complex industries with greatest occupational span. Digressions from this path into mining (light purple), food processing (yellow), or processing of raw materials industries (dark purple), lead to the periphery of the product space. Despite abstracting from many of the detailed connections that are pruned in these graphs, it is interesting to see how the directed nature of pairwise proximities thus highlights well-known and plausible developmental pathways with very different long-run implications for the expected standards of living in a country.

Different development paths on the genotypic product space are also associated with different improvements in standards of living. We illustrate this in Fig. 4. The layout of the network is the same as in Fig. 3. However, colors now reflect the average level of development associated with the industry, measured as the average GDP per capita of the countries that are active in the industry. Industries typically found in countries with low GDP per capita (yellow-orange) are mainly in the bottom-left part of the space, while industries that are more often found in developed countries (red-purple) are mainly positioned in the top-right part of the space.

How different is the genotypic product space from the phenotypic one? To answer this question, Table 2 reports the correlation coefficients among the genotypic proximities of eqs (5) and (6) on the phenotypic proximity of eq. (3). Both genotypic proximity metrics correlate significantly with the phenotypic product space. This suggests that occupational inputs capture important capability requirements that drive a co-exporting patterns. Conversely, it means that conventional phenotypic approaches capture important information regarding the underlying capability structure in terms of occupational inputs. However, the correlations between phenotypic and genotypic proximities are far from perfect, suggesting that both measures provide different types of information that can be exploited in applied work.

Figure 4: Genotypic product space $\tilde{\Gamma}_{pp'}$ and GDP per capita



Notes: Each node represents a 4-digit Naics industry, which is positioned as in Figure 3. However, colors now represent the average *GDP per capita* of countries that are active in the industry: *yellow*=low GDP (first quartile); *orange*=medium-low GDP (second quartile); *red*=medium-high GDP (third quartile); *purple*=high GDP (fourth quartile).

5.2 Genotypic density and export diversification

To test the predictive power of the genotypic approach, we study how countries diversify their exports. To do so, we aggregate our data into 5-year windows: 1992-1996, 1997-2001, 2002-2006, 2007-2011, 2012-2016. Within each 5-year window t, we calculate RCA^t_{cp} using eq. (2). We then define the entry of country c in product p as:

Table 2: Correlation matrix for $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$

	$\Phi_{pp'}$	$\Gamma_{pp'}$	$\tilde{\Gamma}_{pp'}$
$\Phi_{pp'}$	1		
$\Gamma_{pp'}$	0.40	1	
$\tilde{\Gamma}pp'$	0.39	0.87	1

Notes: The table reports on the correlation coefficients for $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$. All correlations are run on $(88^2 - 88)/2 = 3828$ observations. Since the phenotypic product space is symmetric ($\Phi_{pp'} = \Phi_{p'p}$), we also symmetrize the genotypic spaces, calculating for every pair (p, p') the mean of the two directed genotypic distances in (p, p') and (p', p). Tables B.1 and B.2 in Appendix B.1 use the maximum or minimum genotypic distance in (p, p') and (p', p) .

$$y_{cp}^t = \mathbb{1}[(RCA_{cp}^{t+1} - RCA_{cp}^t \geq 0.5)|RCA_{cp}^t < 1]. \tag{10}$$

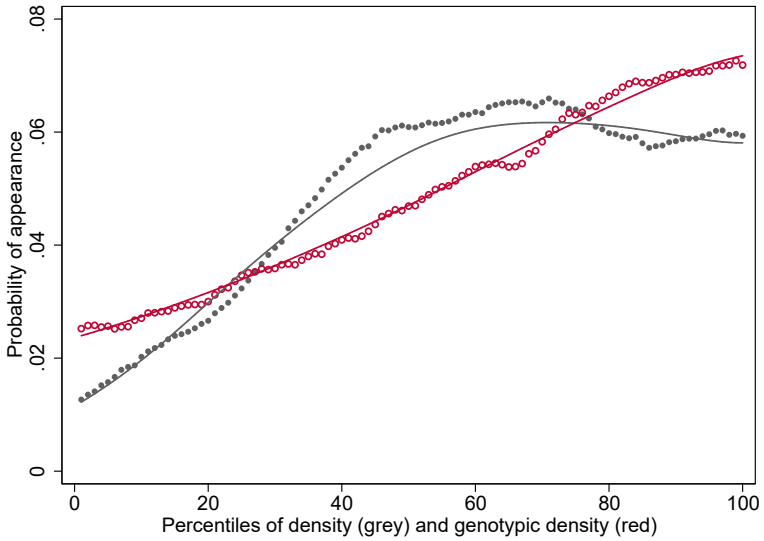
Eq. (10) indicates that a country enters a product, if the RCA makes a jump of 0.5 or higher within a 5-year period. Note that we ignore country-product observations with $RCA_{cp}^t \geq 1$ and thus only consider jumps for countries that do not yet significantly produce the product at time t. In Appendix B.2, we test the robustness of our analysis to different definitions of y_{cp}^t. Next, we calculate for every country-product-time observation both phenotypic (ω_{cp}^t) and genotypic densities (μ_{cp}^t, $\tilde{\mu}_{cp}^t$) as described in Section 3.

Figure 5 shows how the probability of entry changes with density. To do so, we first calculate percentile ranks for our density measures. Next, we create a sliding window across these percentiles that is centered on ω_{cp}^t, respectively $\tilde{\mu}_{cp}^t$ and spans from 20 percentiles below to 20 percentiles above these values. Finally, we calculate in each window the average of y_{cp}^t, the relative frequency at which we observe a country entering a product for the associated density window.

For both measures, the probability of entry runs from 1-2% at low densities to 5-7% at high densities. In this sense, the two measures perform about equally well when it comes to predicting entry events. However, whereas the entry probability rises monotonically with increasing values of genotypic density, it plateaus about mid-way for the phenotypic density. This suggests that the genotypic density ranks products more consistently across its distribution when it comes to the likelihood of countries' diversifying into them. One possible explanation for this is that the genotypic density successfully filters out redundant information from closely related products, whereas the phenotypic density does not.

Table 3 corroborates that phenotypic and genotypic approaches have similar global pre-

Figure 5: Probability of entry and density



Notes: The x-axis represents percentiles of ω_{cp}^t and $\tilde{\mu}_{cp}^t$; the y-axis the probability of appearance for the percentile, which we compute as the average of y_{cp}^t in an interval of ± 20 percentiles around the x-axis value. The plot for phenotypic density (ω_{cp}^t) is drawn with grey dots, while the one for genotypic density ($\tilde{\mu}_{cp}^t$) with red hollow circles. The lines are LOWESS smooths of the plots.

dictive validity. It shows results from the following regression:

$$y_{cp}^t = \beta_1 \log \omega_{cp}^t + \beta_2 \log \tilde{\mu}_{cp}^t + \boldsymbol{\delta}_{cp}^t + \epsilon_{cp}^t, \tag{11}$$

where $\boldsymbol{\delta}_{cp}^t$ is a column-specific vector of fixed effects as indicated in the last row of Table 3, ranging from simple product (p), country (c) and period (t) fixed effects to composite product-period and country-period fixed effects. The first rows of the table show univariate regressions, including only one of the two density measures. The bottom rows show results when both density types enter the regression simultaneously.

Regardless of the included fixed effects, the univariate regressions indicate that phenotypic and genotypic density have similar predictive performance. Furthermore, the multivariate regressions show that the two metrics remain significant also when they are included jointly in the regression. This means that phenotypic and genotypic regressions do not capture exactly the same variation.[20]

[20]To check the robustness of these results, we run a number of regressions (see Appendix B.1), with variations in the definition of the appearance variable y_{cp}^t. Furthermore, Abadie et al. (2023) highlights that—in the absence of a sampling problem—the clustered standard errors could overestimate the true variance. We, thus, report robust standard errors in the main text and cluster-robust ones in the appendix. Especially in the case of the effect of genotypic density, we only find minor differences compared to robust standard errors.

Table 3: Probability of entry and density

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \omega_{cp}^t$	0.016***	0.076***	0.014***	0.007**	0.008***	0.052***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.004)
Adj. R^2	0.01	0.06	0.02	0.05	0.06	0.08
N	37563	37563	37563	37563	37563	37563
$\log \tilde{\mu}_{cp}^t$	0.101***	0.052***	0.124***	0.040***	0.040***	0.061***
	(0.006)	(0.008)	(0.007)	(0.011)	(0.011)	(0.012)
Adj. R^2	0.00	0.05	0.02	0.05	0.06	0.07
N	37570	37570	37570	37570	37570	37570
$\log \omega_{cp}^t$	0.014***	0.076***	0.009***	0.005*	0.006**	0.050***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.004)
$\log \tilde{\mu}_{cp}^t$	0.042***	0.056***	0.065***	0.035***	0.034***	0.042***
	(0.008)	(0.008)	(0.011)	(0.012)	(0.012)	(0.012)
Adj. R^2	0.01	0.07	0.02	0.05	0.06	0.08
N	37563	37563	37563	37563	37563	37563
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports three sets of regressions following Equation 11: the first, including only phenotypic density ($\log \omega_{cp}^t$); the second, only genotypic proximity ($\log \tilde{\mu}_{cp}^t$); the third, both. The dependent variable follows the definition in Equation 10. Robust standard errors in parentheses. Significance is indicated by *(10%), **(5%), and ***(1%).

5.3 Beyond density

A key advantage of the genotypic approach is that it not only allows assessing which industries are close to a country's current export basket, but also which capabilities the country would need to acquire to enter this industry. To illustrate the empirical value of this additional information, we add three explanatory variables to our appearance regression. First, we add req_{cp}^t, the average years of education for the occupations that country c is missing to start making product p, using educational requirements by occupations as provided by the U.S. Bureau of Labor Statistics. Second, we add edu_c^t, the country's average years of education from Barro and Lee (2013). Third, we add the interaction of req_{cp}^t and edu_c^t. This yields the following regression model:

$$y_{cp}^t = \beta_1 \log \tilde{\mu}_{cp}^t + \beta_2 \log req_{cp}^t + \beta_3 \log edu_c^t + \beta_4 \log req_{cp}^t \times \log edu_c^t + \boldsymbol{\delta}_{cp}^t + \epsilon_{cp}^t. \quad (12)$$

Table 4: Capability-enhanced appearance regressions

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \tilde{\mu}_{cp}^t$	0.106***	0.049***	0.136***	0.033***	0.034***	0.056***
	(0.007)	(0.009)	(0.009)	(0.012)	(0.012)	(0.013)
$\log req_{cp}^t$	-0.587***	-0.634***	-0.223**	-0.305***	-0.304***	-0.235**
	(0.084)	(0.086)	(0.100)	(0.102)	(0.102)	(0.102)
$\log edu_c^t$	-0.680***	0.000	-0.352***	-0.436***	-0.362***	
	(0.098)	(0.000)	(0.113)	(0.116)	(0.117)	
$\log req_{cp}^t \times \log edu_c^t$	0.248***	0.261***	0.122***	0.151***	0.151***	0.122***
	(0.037)	(0.038)	(0.043)	(0.044)	(0.044)	(0.044)
Adj. R^2	0.01	0.05	0.03	0.06	0.06	0.07
N	31923	31923	31923	31923	31923	31923
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports the regression described in Equation 12: the regressors are genotypic proximity ($\log \tilde{\mu}_{cp}^t$), the (weighted) average years of education required in missing occupations ($\log req_{cp}^t$), the country's average years of education ($\log edu_c^t$), and the interaction of the latter two terms. The dependent variable follows the definition in Equation 10. Robust standard errors in parentheses. Significance is indicated by *(10%), **(5%), and ***(1%).

Table 4 reports results. We focus here on the interaction between educational requirements for the potential diversification event and the educational endowments of the country. Different columns correspond to models with different types of fixed effects. Across specifications, the interaction effect is positive and highly significant, suggesting a robust complementarity between the nature of the missing capabilities and a country's endowments.[21] This finding strongly resonates with Schetter (2022) who shows how such complementarities provide a flexible microfoundation for the Economic Complexity Index (Hidalgo and Hausmann, 2009) that is in line with key concepts in the related literature.

6 Discussion

The genotypic approach opens up a number of new avenues in economic complexity analysis. Economic development in ECA, but also in the capability approach in evolutionary economics, is often regarded as a process of accumulating capabilities that allow a country to enter into more complex industries. By revealing the capability bases of countries, the genotypic approach allows us to study a country's capability trajectory directly.

[21]Robustness checks are reported in Appendix B.1. Including a control for phenotypic density or changing the measure of education we use does not change results.

To provide a concrete example, consider the economic development process of Vietnam between 1992 and 2016. At the start of this period, Vietnam was active in 14 of industries, such as *Footwear Manufacturing*, *Apparel Manufacturing*, and *Furniture Manufacturing*. Over the next three decades, Vietnam entered a dozen new industries, including *Communications Equipment Manufacturing*, *Audio and Video Equipment Manufacturing*, *Computer and Peripheral Equipment Manufacturing* and *Semiconductor and Other Electronic Component Manufacturing*. According to our analysis, this expansion required acquiring about 40 new capabilities, including the expertise of *Computer Hardware Engineers*, *Materials Scientists*, *Electronics Engineers*, *Electromechanical Equipment Assemblers*, and *Data Communications Analysts*.

However, the potential of the genotypic approach goes well beyond such descriptive analysis. Below, we sketch how the genotypic approach can be used in policy making and to develop a new research agenda within ECA.

6.1 Implications for policy

ECA has contributed to policy frameworks for country-level and regional economic development. It has been applied by policy makers in multilateral organizations such as the World Bank (Bank, 2019) and the European Union, where it offers analytic insights for smart specialization (Boschma et al., 2021; Diodato et al., 2023b), one of the world's largest place-based policy actions. Our genotypic analysis can augment such policy frameworks in several ways.

The genotypic approach goes beyond measures of proximity or density and complexity rankings, offering insights into actual capability requirements and capability endowments. In practical terms, such insights can be leveraged to determine the feasibility of diversification paths in new ways that directly ask which capabilities are missing and which actors or economies may be the best sources to acquire them from. This information moves policy prescriptions from targeting specific products to investing in specific capabilities. This, in turn, limits the risks of capture by vested interests of actors that operate in specific product markets and instead turns the focus to targeted provision of public goods. In this sense, the genotypic approach may also support Lin's (2011) NSE framework by helping identify which hard and soft infrastructure a country should aim to develop.

Furthermore, although the genotypic approach does not have a clear edge in predicting diversification paths over phenotypic approaches, the linear mapping between genotypic density and entry probabilities of Fig. 5 suggests that it does more accurately map the

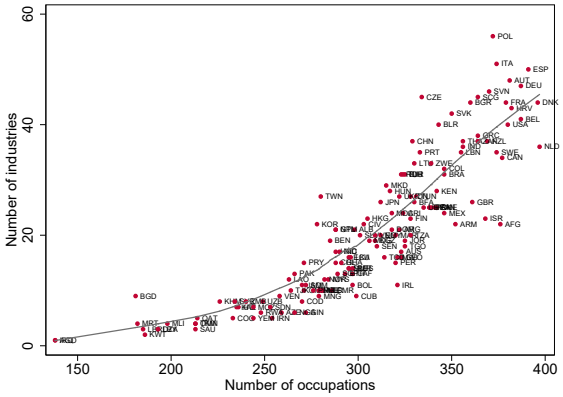
relative risk ratios of diversification over a broader range of densities. This is useful because it can help a country better assess trade-offs between feasibility and desirability of entering a new product at different points along the density spectrum.

The genotypic approach also has implications for long-term economic development. So far, we have focused on the ability of the product space to predict diversification patterns in the short- to medium-run. However, the directed nature of the genotypic product space reveals how different choices can lead to different long-run developmental options, as well as suggest good ways to sequence the acquisition of capabilities along such paths. Moreover, it draws attention to a number of potential poverty traps (Hidalgo et al., 2007; Hausmann and Hidalgo, 2011; Tacchella et al., 2016; Diodato et al., 2022). Explicitly accounting for the capabilities allows the genotypic approach to provide additional insights that cannot be derived at the phenotypic level.

For instance, the genotypic approach allows to test the empirical relevance of the so-called *quiescence trap* (Hausmann and Hidalgo, 2011; Tacchella et al., 2016). The quiescence trap refers to the hypothesis that the combinatorial dynamics that ECA assumes for economic production may lock a country into periods of developmental stasis (Hausmann and Hidalgo, 2011). This follows directly from the production framework assumed in ECA, according to which countries only produce the products for which they have all required capabilities. As countries acquire new capabilities, they can combine them with their existing capabilities. This leads to new capability combinations, some of which are associated with viable products. However, the value of a new capability depends on the number of capabilities a country already has: one extra capability leads to many more new combinations in highly developed economies that already dispose of many capabilities than in less developed countries with only few existing capabilities.

The genotypic approach allows taking these ideas to the data. To do so, Fig. 6 plots for each country its number of capabilities on the horizontal and its number of industries on the vertical axis. The relationship between the number of capabilities a country has and the number of industries in which it is active exhibits the hypothesized convexity needed for quiescence traps. In particular, the graph shows that at low levels of diversification, i.e., among countries with ~ 270 capabilities or less, industrial diversification grows only little with the number of capabilities. For instance Angola has 138 capabilities and 1 industry, while Qatar has 214 capabilities and 5 industries. As we move to countries with more capabilities, the number of industries rises rapidly: Albania has 298 capabilities but already 21 industries, while Germany has 387 capabilities and 47 industries. This suggests

Figure 6: Quiescence trap



Notes: Scatter plot of the number of industries (M_c) against number of revealed capabilities (occupations, C_c) per country. The graph refers to the year 2002 and uses an RCA threshold of 1. The trend line is a LOWESS smooth.

that the number of industries that a country can develop only rises rapidly at later stages of development.

The genotypic approach can also deepen our understanding of other poverty traps. For instance, Hidalgo et al. (2007) already showed that countries specialized in products at the periphery of the product space tend to have relatively few diversification opportunities, suggesting the existence of a *periphery trap*. In Appendix C.1, we discuss how the genotypic product space can be used to simulate diversification opportunities and show that the capability requirements of *Apparel accessories* are more conducive of development than those of *Metal ore mining*, for example.

Finally, a related, but distinct, poverty trap is discussed in Appendix C.2. Here, we argue that diversification might "congest" capabilities: entering new industries increases the opportunity cost for hiring labor for further diversification moves. Consequently, diversifying into new industries has two opposing effects: (i) it allows the country to accumulate new capabilities, thus making it easier to further diversify; (ii) it raises the cost of existing capabilities, making it more difficult to further diversify. Depending on the balance of these two effects, the order of entry into new industries matters. Although this *entry trap* may, in principle, also be studied at the phenotypic level, the genotypic approach's attention to opportunity costs of the non-tradeable inputs of production offers a more natural framework to do so (see also Diodato et al., 2022).

6.2 A new research agenda

The analysis so far has provided a sketch of what a genotypic approach to economic complexity may look like and how it can be used in development research and policy. To do so, we studied the framework in a highly stylized setting and there are many ways in which this approach can be refined and extended. Moreover, going down to the level of capabilities opens up a range of new questions and opportunities for further analysis.

First, our focus on human capital inputs strongly has limited the set of capabilities we have considered. The advantage of doing so is that it kept the exposition compact. Moreover, human capital related capabilities are likely to fulfill important characteristics that are assumed about capabilities in the stylized model of production we use: human capital is valuable, relatively specific, mostly non-ubiquitous and hard to move or access from outside the country where it resides. Moreover, many other types of capabilities have components that are embedded in the skills of workers. For instance, physical equipment and technological expertise require workers that know how to use this equipment and apply the expertise. However, future research could focus on various other types of capabilities that can be mapped using the methodology described in this paper. For example, nontraded intermediates, such as specialized business services, may fit our definition of capabilities as well. Information on which industries rely on which business services is readily available in the supply and use tables used in input-output analysis, suggesting that such information can be added relatively easily to the $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ matrices. Another example is the technological areas in which industries conduct R&D as revealed in patent data.

Second, even if we limit ourselves to human-capital related capabilities, our analysis can be augmented. After all, occupations are themselves bundles of tasks and some pairs of occupations are more similar to one another than others (Gathmann and Schönberg, 2010; Neffke et al., 2024). Therefore, although we have avoided double-counting occupations, we may still double-count capabilities in terms of the underlying skills, knowledge and abilities of workers in these occupations. Datasets that describe the content of occupations such as O*NET in the US may help remedy such problems, expressing capability requirements of industries in terms of the skills they rely on through the human capital of their workers. This type of analysis may also help connect the genotypic approach in ECA to the task-based approach in labor economics (Acemoglu and Autor, 2011), as well as to research about the future of work (e.g., Alabdulkareem et al., 2018).

Third, an interesting set of questions arises from relaxing the assumption of a universal

capability requirements matrix. To a first approximation, this may be justified: although car firms may differ in their exact technologies and human capital, it is plausible that capability requirements of car industries in different countries are more similar to one another than they are to the capability mix required by mining companies. Nevertheless, the mix of occupations that a given industry employs may differ between industrialized and less developed economies. In this case, the $\boldsymbol{P}$ matrix differ across (groups of) countries. Similarly, capability requirements may change. Such variations of the $\boldsymbol{P}$ matrix can be easily accommodated. For instance, we may use different versions of $\boldsymbol{P}$ for advanced and developing economies. Furthermore, future research could explore how to extend our framework to allow for countries to choose between different production technologies.

Fourth, we have tested the genotypic analysis in ECA's original context of international trade. However, economic complexity has had a large influence on the field of economic geography and in particular of evolutionary economic geography. Therefore, studying the industrial diversification dynamics of cities and regions could offer a particularly promising alternative application.

Fifth, although the acquisition of new capabilities has played a prominent role in our paper, we have remained silent on *how* countries develop new capabilities. An expanding literature has argued that capabilities often diffuse from places where they are already well established. This literature has identified several channels for such capability diffusion, from migration (e.g., Bahar and Rapoport, 2018; Diodato et al., 2023a), to FDI (e.g., Crescenzi et al., 2022) and business travel (Coscia et al., 2020). By revealing changes in the capability mix of countries, the genotypic approach offers new, more direct, ways of analyzing these channels and their importance.

Sixth, our work does not focus on maximizing predictive validity, as in Tacchella et al. (2023). However, the genotypic predictions can be improved by fine-tuning e.g. when to treat a capability as present in a country or how different capabilities are weighted. For instance, one can use information on the value or difficulty of acquiring a capability. It may also be possible to calibrate these weights such that they maximize the predictive validity of genotypic density metrics in diversification dynamics. Such an analysis would provide valuable information on how important different capabilities are.

Seventh, although we have not fully pursued this, the genotypic approach also directly suggests genotypic complexity metrics. This could be as simple as counting the capabilities of a country. Comparing the predictive validity of such genotypic complexity measures to their phenotypic counterparts in predictions of GDP per capita growth may shed further

light on the relative merits of each approach. Incidentally, this exercise may also offer a different way to tune the aforementioned capability weights.

Finally, a number of ECA-inspired combinatorial models of economic growth have been proposed (Hausmann and Hidalgo, 2011; Fink et al., 2017; van Dam and Frenken, 2020). By making capabilities measurable, the genotypic approach offers ways to test these models more directly.

In sum, although the genotypic approach to ECA we have sketched has shown promising results, much work remains. To facilitate this work, we make the underlying Python code as well as the estimated capability bases of countries and their changes over time available for download.

7 Conclusions

Economic complexity analysis has advanced our knowledge of diversification and structural transformation processes. The strength of this literature is that it showed how detailed, highly disaggregated information on an economy's industrial structure can be analyzed to map stylized development trajectories that are predictive of future diversification, without resorting to a small set of factors of production or coarse stages of development. So far, this literature has used the notion of capabilities primarily as a narrative that underlies methods to derive product similarities and complexity rankings. Here, we propose that by taking the capabilities narrative more literally, we can arrive at more informative descriptions of countries' capability bases. This offers new ways to analyze the costs countries face when trying to enter new products and how they can achieve this. We have termed this approach genotypic, using capabilities as akin to encoding a DNA of products and countries, where different combinations of capabilities result in different products. This stands in contrast to more traditional economic complexity analyses, which we have termed phenotypic, because they group and rank products and countries based on observed outcomes alone, without direct reference to the underlying capability structure.

Although our genotypic approach uses an arguably crude approximation of how economies produce products, it emphasizes important constraints and opportunities in economic development that have first-order consequences for how economies move along their development trajectories. Moreover, the proposed genotypic approach complements conventional product space analyses and helps deepen our understanding of underlying mechanisms in

various ways. First and foremost, directly building on the underlying capability structure opens up the black box of what drives co-location and related diversification. Second, the focus on capabilities allows deriving proximity and density metrics that are, in principle, consistent with the underlying drivers of technological similarity. This is not, in general, the case for phenotypic analyses—see Appendix D.1. Third, the genotypic approach suggests that the proximity between products is *directed*: diversifying from textiles to airplanes is not the same as from airplanes to textiles. We have shown that this asymmetry has important short-run and long-run consequences for economic development. Lastly, building on the capability structure opens up new avenues that are impossible to explore at the phenotypic level: Co-location patterns can indicate which products are likely to be in a country's adjacent possible, but they cannot tell us what it would take for a country to actually start making these products. The genotypic approach allows filling this gap, opening up new avenues for research and approaches to policy making.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35.

Abramovitz, M. (1956). Resource and output trends in the united states since 1870. In *Resource and output trends in the United States since 1870*, pages 1–23. NBER.

Abramovitz, M. (1986). Catching up, forging ahead, and falling behind. *The journal of economic history*, 46(2):385–406.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4, part B, chapter 12, pages 1043–1171. Elsevier, Amsterdam.

Alabdulkareem, A., Frank, M. R., Sun, L., AlShebli, B., Hidalgo, C., and Rahwan, I. (2018). Unpacking the polarization of workplace skills. *Science advances*, 4(7):eaao6030.

Aldrich, H. E., Hodgson, G. M., Hull, D. L., Knudsen, T., Mokyr, J., and Vanberg, V. J. (2008). In defence of generalized darwinism. *Journal of evolutionary economics*, 18:577–596.

Archibugi, D. and Coco, A. (2005). Measuring technological capabilities at the country level: A survey and a menu for choice. *Research policy*, 34(2):175–194.

Armington, P. S. (1969). A theory of demand for products distinguished by place of production. *Staff Papers - International Monetary Fund*, 16(1):159.

Atkin, D., Costinot, A., and Fukui, M. (2021). Globalization and the ladder of development: Pushed to the top or held at the bottom? Working Paper 29500, NBER. Series: Working Paper Series.

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

Bahar, D. and Rapoport, H. (2018). Migration, knowledge diffusion and the comparative advantage of nations. *The Economic Journal*, 128(612):F273–F305.

Bahar, D., Rosenow, S., Stein, E., and Wagner, R. (2019). Export take-offs and acceleration: Unpacking cross-sector linkages in the evolution of comparative advantage. *World Development*, 117:48–60.

Balassa, B. (1965). Trade liberalisation and 'revealed' comparative advantage. *The Manchester School*, 33(2):99–123.

Balland, P.-A., Boschma, R., Crespo, J., and Rigby, D. L. (2018). Smart specialization policy in the european union: relatedness, knowledge complexity and regional diversification. *Regional studies*.

Bank, W. (2019). *Chad Growth and DiversificationLeveraging Export Diversification to Foster Growth: Leveraging Export Diversification to Foster Growth*. World Bank.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management*, 17(1):99–120.

Barro, R. J. and Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of development economics*, 104:184–198.

Boschma, R. et al. (2021). *Designing Smart Specialization Policy: relatedness, unrelatedness, or what?* Utrecht University, Human Geography and Planning.

Boschma, R. A. and Frenken, K. (2006). Why is economic geography not an evolutionary science? Towards an evolutionary economic geography. *Journal of economic geography*, 6(3):273–302.

Buera, F. J., Kaboski, J. P., Rogerson, R., and Vizcaino, J. I. (2022). Skill-biased structural change. *Review of Economic Studies*, 89(2):592–625.

Bustos, S., Gomez, C., Hausmann, R., and Hidalgo, C. A. (2012). The dynamics of nestedness predicts the evolution of industrial ecosystems. *PLOS ONE*, 7(11):e49393.

Coscia, M., Neffke, F. M., and Hausmann, R. (2020). Knowledge diffusion in the network of international business travel. *Nature Human Behaviour*, 4(10):1011–1020.

Costinot, A. (2009). An elementary theory of comparative advantage. *Econometrica*, 77(4):1165–1192.

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

Crescenzi, R., Dyèvre, A., and Neffke, F. (2022). Innovation catalysts: how multinationals reshape the global geography of innovation. *Economic Geography*, 98(3):199–227.

Diodato, D., Hausmann, R., and Neffke, F. (2023a). The impact of return migration on employment and wages in mexican cities. *Journal of Urban Economics*, 135:103557.

Diodato, D., Hausmann, R., and Schetter, U. (2022). A simple theory of economic development at the extensive industry margin. *HKS Working Paper No. RWP22-016*.

Diodato, D., Napolitano, L., Pugliese, E., and Tacchella, A. (2023b). Economic complexity for regional industrial strategies. *JRC Science for Policy Brief - Industrial Innovation & Dynamics Series.*, No. JRC136443. European Commission, Joint Research Centre.

Diodato, D., Neffke, F., and O'Clery, N. (2018). Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26.

Ellison, G., Glaeser, E., and Kerr, W. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Fagerberg, J. and Srholec, M. (2008). National innovation systems, capabilities and economic development. *Research policy*, 37(9):1417–1435.

Fagerberg, J., Srholec, M., and Verspagen, B. (2010). Innovation and economic development. In *Handbook of the Economics of Innovation*, volume 2, pages 833–872. Elsevier.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the penn world table. *American economic review*, 105(10):3150–3182.

Fink, T., Reeves, M., Palma, R., and Farr, R. (2017). Serendipity and strategy in rapid innovation. *Nature Communications*, 8(1):2002.

Foellmi, R. and Zweimüller, J. (2008). Structural change, Engel's consumption cycles and Kaldor's facts of economic growth. *Journal of Monetary Economics*, 55(7):1317–1328.

Frenken, K. and Boschma, R. A. (2007). A theoretical framework for evolutionary economic geography: industrial dynamics and urban growth as a branching process. *Journal of economic geography*, 7(5):635–649.

Gathmann, C. and Schönberg, U. (2010). How general is human capital? a task-based approach. *Journal of Labor Economics*, 28(1):1–49.

Gersbach, H., Schetter, U., and Schmassmann, S. (2023). From local to global: A theory of public basic research in a globalized world. *European Economic Review*, 160:104530.

Hausmann, R. and Hidalgo, C. A. (2011). The network structure of economic output. *Journal of economic growth*, 16:309–342.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., and Yildirim, M. A. (2011). *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. https://atlas.media.mit.edu/atlas/.

Hausmann, R., Hwang, J., and Rodrik, D. (2007). What you export matters. *Journal of Economic Growth*, 12(1):1–25.

Hausmann, R. and Klinger, B. (2006). Structural transformation and patterns of comparative advantage in the product space. Working Paper 128, CID at Harvard University.

Hausmann, R. and Rodrik, D. (2003). Economic development as self-discovery. *Journal of Development Economics*, 72(2):603–633.

Hidalgo, C. A. (2023). The policy implications of economic complexity. *Research Policy*, 52(9):104863.

Hidalgo, C. A., Balland, P.-A., Boschma, R., Delgado, M., Feldman, M., Frenken, K., Glaeser, E., He, C., Kogler, D. F., Morrison, A., Neffke, F., Rigby, D., Stern, S., Zheng, S., and Zhu, S. (2018). The principle of relatedness. In Morales, A. J., Gershenson, C., Braha, D., Minai, A. A., and Bar-Yam, Y., editors, *Unifying Themes in Complex Systems IX*, Springer Proceedings in Complexity, pages 451–457, Cham. Springer International Publishing.

Hidalgo, C. A. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575.

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

Hirschman, A. O. (1958). The strategy of economic development. *(No Title)*.

Kim, L. (1980). Stages of development of industrial technology in a developing country: a model. *Research policy*, 9(3):254–277.

Kogler, D. F., Rigby, D. L., and Tucker, I. (2015). Mapping knowledge space and technological relatedness in us cities. In *Global and Regional Dynamics in Knowledge Flows and Innovation*, pages 58–75. Routledge.

Kongsamut, P., Rebelo, S., and Xie, D. (2001). Beyond balanced growth. *Review of Economic Studies*, 68(4):869–882.

Krugman, P. (1985). A 'technology gap' model of international trade. In Jungenfelt, K. and Hague, D., editors, *Structural Adjustment in Developed Open Economies*, pages 35–61. Palgrave Macmillan UK, London.

Kuznets, S. (1957). Quantitative aspects of the economic growth of nations: II. Industrial distribution of national product and labor force. *Economic Development and Cultural Change*, 5(4):1–111.

Lall, S. (1992). Technological capabilities and industrialization. *World development*, 20(2):165–186.

Li, Y. and Neffke, F. (2023). Evaluating the principle of relatedness: Estimation, drivers and implications for policy. *CID Research Fellows and Graduate Student Working Paper Series*.

Lin, J. Y. (2011). New structural economics: A framework for rethinking development. *The World Bank Research Observer*, 26(2):193–221.

Lucas, R. E. (1993). Making a miracle. *Econometrica*, 61(2):251–272.

Matsuyama, K. (1992). Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory*, 58(2):317–334.

Matsuyama, K. (2019). Engel's law in the global economy: Demand-induced patterns of structural change, innovation, and trade. *Econometrica*, 87(2):497–528.

Neffke, F., Hartog, M., Boschma, R., and Henning, M. (2018). Agents of structural change: The role of firms and entrepreneurs in regional diversification. *Economic Geography*, 94(1):23–48.

Neffke, F., Henning, M., and Boschma, R. (2011). How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic Geography*, 87(3):237–265.

Neffke, F., Nedelkoska, L., and Wiederhold, S. (2024). Skill mismatch and the costs of job displacement. *Research Policy*, 53(2):104933.

Nelson, R. R. and Winter, S. (1982). *An evolutionary theory of economic change*. harvard university press.

O'Clery, N., Yildirim, M. A., and Hausmann, R. (2021). Productive ecosystems and the arrow of development. *Nature Communications*, 12(1):1479.

Pierce, J. R. and Schott, P. K. (2009). A concordance between ten-digit U.S. harmonized system codes and SIC/NAICS product classes and industries. Working Paper 15548, NBER.

Polanyi, M. (1962). The republic of science. *Minerva*, 1(1):54–73.

Prebisch, R. (1962). The economic development of latin america and its principal problems. *Economic Bulletin for Latin America*.

Pugliese, E., Cimini, G., Patelli, A., Zaccaria, A., Pietronero, L., and Gabrielli, A. (2019). Unfolding the innovation system for the development of countries: coevolution of science, technology and production. *Scientific reports*, 9(1):16440.

Schetter, U. (2020). Quality differentiation, comparative advantage, and international specialization across products. *CID Research Fellow and Graduate Student Working Paper*, (126). http://dx.doi.org/10.2139/ssrn.3091581.

Schetter, U. (2022). A measure of countries? distance to frontier based on comparative advantage. *CID Research Fellow and Graduate Student Working Paper*, (135). http://dx.doi.org/10.2139/ssrn.4227848.

Steijn, M. P., Koster, H. R., and Van Oort, F. G. (2022). The dynamics of industry agglomeration: Evidence from 44 years of coagglomeration patterns. *Journal of Urban Economics*, 130:103456.

Sutton, J. and Trefler, D. (2016). Capabilities, wealth, and trade. *Journal of Political Economy*, 124(3):826–878.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific Reports*, 2:723. DOI: 10.1038/srep00723.

Tacchella, A., Di Clemente, R., Gabrielli, A., and Pietronero, L. (2016). The build-up of diversity in complex ecosystems. *arXiv preprint arXiv:1609.03617*.

Tacchella, A., Zaccaria, A., Miccheli, M., and Pietronero, L. (2023). Relatedness in the era of machine learning. *Chaos, Solitons & Fractals*, 176:114071.

Teece, D. J., Rumelt, R., Dosi, G., and Winter, S. (1994). Understanding corporate coherence: Theory and evidence. *Journal of economic behavior & organization*, 23(1):1–30.

Uy, T., Yi, K.-M., and Zhang, J. (2013). Structural change in an open economy. *Journal of Monetary Economics*, 60(6):667–682.

van Dam, A. and Frenken, K. (2020). Variety, complexity and economic development. *Research Policy*, page 103949.

Appendices

A Supplementary descriptives

A.1 Rankings of countries, products, and occupations

Table A.1: Countries with fewer industries

# Ind.	ISO3	Country name
1	AGO	Angola
1	IRQ	Iraq
2	KWT	Kuwait
3	DZA	Algeria
3	LBR	Liberia
3	LBY	Libya
3	SAU	Saudi Arabia
4	MLI	Mali
4	MRT	Mauritania
4	OMN	Oman
4	TKM	Turkmenistan
5	COG	Congo
5	IRN	Iran
5	QAT	Qatar
5	YEM	Yemen

Notes: The table displays the bottom ranking of M_c (country diversity $= \sum_p M_{cp}$).

Table A.2: Countries with most industries

# Ind.	ISO3	Country name
56	POL	Poland
51	ITA	Italy
50	ESP	Spain
48	AUT	Austria
47	DEU	Germany
46	SVN	Slovenia
45	CZE	Czechia
45	SCG	Serbia and Montenegro
44	BGR	Bulgaria
44	DNK	Denmark
44	FRA	France
43	HRV	Croatia
42	SVK	Slovakia
41	BEL	Belgium-Luxembourg
40	BLR	Belarus

Notes: The table displays the top ranking of M_c (country diversity $= \sum_p M_{cp}$).

Table A.3: Industries in fewer countries

# Countries	Naics	Industry name
11	3333	Commercial and Service Industry Machinery
11	3327	Screws, Nuts, and Bolts
12	3364	Aerospace Products and Parts
13	3339	Other General Purpose Machinery
13	3344	Semiconductors and Other Electronic Components
15	3336	Engine, Turbine, and Power Transmission Equipment
15	3332	Industrial Machinery
15	3345	Navigational and Electromedical Instruments
16	3343	Audio and Video Equipment
16	3342	Communications Equipment
16	3351	Electric Lighting Equipment
16	3369	Other Transportation Equipment
16	3346	Magnetic and Optical Media
17	3341	Computer and Peripheral Equipment
17	3325	Hardware

Notes: The table displays the bottom ranking of M_p (industry ubiquity $= \sum_c M_{cp}$).

Table A.4: Industries in most countries

# Countries	Naics	Industry name
67	3119	Other Food
64	3113	Sugar and Confectionery
58	3116	Animal Slaughtering and Processing
58	3273	Cement and Concrete Products
58	3161	Leather and Hide Tanning
57	3114	Fruit and Vegetable Preserving
56	3112	Grain and Oilseed Milling
56	2123	Nonmetallic Mineral Mining
55	3149	Other Textile Product Mills
55	3241	Petroleum and Coal Products
54	3152	Cut and Sew Apparel
54	3253	Pesticide, Fertilizer, Agricultural Chemicals
53	3211	Sawmills and Wood Preservation
52	3219	Other Wood Products
49	3122	Tobacco

Notes: The table displays the top ranking of M_p (industry ubiquity $= \sum_c M_{cp}$).

Table A.5: Industries requiring fewest occupations

# Occ.	Naics	Industry name
39	3161	Leather and Hide Tanning
48	3162	Footwear
51	3169	Other Leather Products
59	3361	Motor Vehicle
62	3159	Apparel Accessories
62	3365	Railroad Rolling Stock
63	3122	Tobacco
68	3274	Lime and Gypsum Products
68	3117	Seafood Preparation and Packaging
72	3151	Apparel Knitting Mills
74	3131	Fiber, Yarn, and Thread Mills
76	2122	Metal Ore Mining
89	2121	Coal Mining
92	3325	Hardware
92	3379	Other Furniture Related Products

Notes: The table displays the bottom ranking of P_p (industry span $= \sum_a P_{pa}$).

Table A.6: Industries requiring most occupations

# Occ.	Naics	Industry name
204	3345	Navigational and Electromedical Instruments
199	3391	Medical Equipment and Supplies
196	3261	Plastics Products
191	3329	Other Fabricated Metal Products
186	3363	Motor Vehicle Parts
180	3339	Other General Purpose Machinery
176	3254	Pharmaceutical and Medicine
175	3344	Semiconductors and Other Electronic Components
172	3364	Aerospace Products and Parts
171	3323	Architectural and Structural Metals
167	3231	Printing and Support Activities
162	3219	Other Wood Products
161	3331	Agriculture, Construction, and Mining Machinery
159	3251	Basic Chemicals
157	3333	Commercial and Service Industry Machinery

Notes: The table displays the top ranking of P_p (industry span $= \sum_a P_{pa}$).

Table A.7: Occupations required by fewest industries

# Ind.	Soc	Occupation name
1	53-2021	Air Traffic Controllers
1	53-2022	Airfield Operations Specialists
1	19-2021	Atmospheric and Space Scientists
1	29-1121	Audiologists
1	51-5012	Bookbinders
1	27-4012	Broadcast Technicians
1	27-4031	Camera Operators, Television, Video, and Motion Picture
1	49-9061	Camera and Photographic Equipment Repairers
1	35-1011	Chefs and Head Cooks
1	39-9011	Child Care Workers
1	35-2014	Cooks, Restaurant
1	31-9091	Dental Assistants
1	29-2021	Dental Hygienists
1	51-9081	Dental Laboratory Technicians
1	47-5011	Derrick Operators, Oil and Gas

Notes: The table displays the bottom ranking of P_a (capability generality $= \sum_p P_{pa}$).

Table A.8: Occupations required by most industries

# Ind.	Soc	Occupation name
88	43-3031	Bookkeeping, Accounting, and Auditing Clerks
88	51-1011	First-Line Supervisors/Managers of Production and Operating Workers
88	43-9061	Office Clerks, General
88	43-5071	Shipping, Receiving, and Traffic Clerks
87	13-2011	Accountants and Auditors
87	43-6011	Executive Secretaries and Administrative Assistants
87	11-3031	Financial Managers
87	43-1011	First-Line Supervisors/Managers of Office and Administrative Support Workers
87	11-1021	General and Operations Managers
87	51-9198	Helpers–Production Workers
87	51-9061	Inspectors, Testers, Sorters, Samplers, and Weighers
87	37-2011	Janitors and Cleaners, Except Maids and Housekeeping Cleaners
87	43-5061	Production, Planning, and Expediting Clerks
87	41-4012	Sales Representatives, Wholesale and Manufacturing, Except Technical Products
87	43-6014	Secretaries, Except Legal, Medical, and Executive

Notes: The table displays the top ranking of P_a (capability generality $= \sum_p P_{pa}$).

Table A.9: Countries endowed with fewest occupations

# Occ.	ISO3	Country name
138	AGO	Angola
138	IRQ	Iraq
181	BGD	Bangladesh
182	MRT	Mauritania
185	LBR	Liberia
186	KWT	Kuwait
193	DZA	Algeria
193	LBY	Libya
198	MLI	Mali
213	OMN	Oman
213	SAU	Saudi Arabia
213	TKM	Turkmenistan
214	QAT	Qatar
226	KHM	Cambodia
233	COG	Congo

Notes: The table displays the bottom ranking of C_c (country completeness $= \sum_a C_{ca}$).

Table A.10: Countries endowed with most occupations

# Occ.	ISO3	Country name
397	NLD	Netherlands
396	DNK	Denmark
391	ESP	Spain
387	BEL	Belgium-Luxembourg
387	DEU	Germany
382	HRV	Croatia
381	AUT	Austria
380	USA	USA
379	FRA	France
377	CAN	Canada
376	AFG	Afghanistan
374	ITA	Italy
374	SWE	Sweden
372	POL	Poland
370	SVN	Slovenia

Notes: The table displays the top ranking of C_c (country completeness $= \sum_a C_{ca}$).

Table A.11: Occupations present in fewest countries

# Countries	Soc	Country name
12	53-2021	Air Traffic Controllers
12	53-2022	Airfield Operations Specialists
12	27-4012	Broadcast Technicians
12	41-3041	Travel Agents
13	47-4021	Elevator Installers and Repairers
15	19-2021	Atmospheric and Space Scientists
15	29-1121	Audiologists
15	49-9061	Camera and Photographic Equipment Repairers
15	15-2091	Mathematical Technicians
15	19-2012	Physicists
15	53-6051	Transportation Inspectors
15	49-9064	Watch Repairers
16	49-2097	Electronic Home Entertainment Equipment Installers and Repairers
16	49-3052	Motorcycle Mechanics
16	27-2012	Producers and Directors

Notes: The table displays the bottom ranking of C_a (capability dispersion $= \sum_c C_{ca}$).

Table A.12: Occupations present in most countries

# Countries	Soc	Country name
140	13-2011	Accountants and Auditors
140	11-3011	Administrative Services Managers
140	49-3023	Automotive Service Technicians and Mechanics
140	43-3021	Billing and Posting Clerks and Machine Operators
140	43-3031	Bookkeeping, Accounting, and Auditing Clerks
140	49-3031	Bus and Truck Mechanics and Diesel Engine Specialists
140	17-2041	Chemical Engineers
140	19-2031	Chemists
140	11-1011	Chief Executives
140	53-7061	Cleaners of Vehicles and Equipment
140	13-1072	Compensation, Benefits, and Job Analysis Specialists
140	43-9011	Computer Operators
140	15-1021	Computer Programmers
140	15-1041	Computer Support Specialists
140	15-1051	Computer Systems Analysts

Notes: The table displays the top ranking of C_a (capability dispersion $= \sum_c C_{ca}$).

B Robustness of Empirics

B.1 Robustness of product space regressions

Table B.1: Cross-regressions of $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$ (max distance within pair)

	$\Phi_{pp'}$	$\Gamma_{pp'}$	$\tilde{\Gamma}_{pp'}$
$\Phi_{pp'}$	1		
$\Gamma_{pp'}$	0.37	1	
$\tilde{\Gamma}pp'$	0.38	0.64	1

Notes: The table reports on the correlation coefficients for $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$. All correlations are run on $(88^2 - 88)/2 = 3828$ observations. To compute the values of Γ and $\tilde{\Gamma}$, we take the maximum genotypic distance between pp' and $p'p$.

Table B.2: Cross-regressions of $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$ (min distance within pair)

	$\Phi_{pp'}$	$\Gamma_{pp'}$	$\tilde{\Gamma}_{pp'}$
$\Phi_{pp'}$	1		
$\Gamma_{pp'}$	0.26	1	
$\tilde{\Gamma}pp'$	0.35	0.75	1

Notes: The table reports on the correlation coefficients for $\Phi_{pp'}$, $\Gamma_{pp'}$, and $\tilde{\Gamma}_{pp'}$. All correlations are run on $(88^2 - 88)/2 = 3828$ observations. To compute the values of Γ and $\tilde{\Gamma}$, we take the minimum genotypic distance between pp' and $p'p$.

B.2 Robustness of appearance regressions

Table B.3: Different definition of appearance (1)

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \omega_{cp}^t$	0.015***	0.065***	0.012***	0.009***	0.010***	0.047***
	(0.001)	(0.003)	(0.001)	(0.002)	(0.002)	(0.004)
Adj. R^2	0.01	0.05	0.02	0.04	0.04	0.06
N	37563	37563	37563	37563	37563	37563
$\log \tilde{\mu}_{cp}^t$	0.082***	0.040***	0.103***	0.034***	0.034***	0.048***
	(0.006)	(0.007)	(0.006)	(0.010)	(0.010)	(0.011)
Adj. R^2	0.00	0.04	0.02	0.04	0.04	0.06
N	37570	37570	37570	37570	37570	37570
$\log \omega_{cp}^t$	0.013***	0.065***	0.010***	0.008***	0.008***	0.045***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.004)
$\log \tilde{\mu}_{cp}^t$	0.026***	0.043***	0.042***	0.025**	0.025**	0.031***
	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)	(0.011)
Adj. R^2	0.01	0.05	0.02	0.04	0.04	0.06
N	37563	37563	37563	37563	37563	37563
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports three sets of regressions following Equation 11: the first, including only phenotypic density ($\log \omega_{cp}^t$); the second, only genotypic proximity ($\log \tilde{\mu}_{cp}^t$); the third, both. The dependent variable is defined as $y_{cp}^t = \mathbb{1}[(RCA_{cp}^{t+1} - RCA_{cp}^t \geq 0.5) \cap RCA_{cp}^{t+1} > 1 | RCA_{cp}^t < 1]$. Robust standard errors in parentheses. Significance is indicated by *(10%), **(5%), and ***(1%).

Table B.4: Different definition of appearance (2)

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \omega_{cp}^t$	0.028***	0.091***	0.026***	0.022***	0.022***	0.080***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.005)
Adj. R^2	0.02	0.05	0.03	0.04	0.05	0.06
N	37563	37563	37563	37563	37563	37563
$\log \tilde{\mu}_{cp}^t$	0.144***	0.077***	0.173***	0.052***	0.052***	0.066***
	(0.007)	(0.008)	(0.008)	(0.012)	(0.012)	(0.013)
Adj. R^2	0.01	0.04	0.02	0.04	0.04	0.06
N	37570	37570	37570	37570	37570	37570
$\log \omega_{cp}^t$	0.027***	0.092***	0.026***	0.020***	0.021***	0.079***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.005)
$\log \tilde{\mu}_{cp}^t$	0.029***	0.081***	0.008	0.029**	0.028**	0.037***
	(0.007)	(0.008)	(0.012)	(0.012)	(0.012)	(0.013)
Adj. R^2	0.02	0.05	0.03	0.04	0.05	0.06
N	37563	37563	37563	37563	37563	37563
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports three sets of regressions following Equation 11: the first, including only phenotypic density (log ω_{cp}^t); the second, only genotypic proximity (log $\tilde{\mu}_{cp}^t$); the third, both. The dependent variable is defined as $y_{cp}^t = \mathbb{1}[RCA_{cp}^{t+1} > 1 | RCA_{cp}^t < 1]$. Robust standard errors in parentheses.Significance is indicated by *(10%), **(5%), and ***(1%).

Table B.5: Two-way clustered standard errors.

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \omega_{cp}^t$	0.016***	0.076***	0.014***	0.007	0.008	0.052***
	(0.002)	(0.012)	(0.002)	(0.006)	(0.006)	(0.014)
Adj. R^2	0.01	0.06	0.02	0.05	0.06	0.08
N	37563	37563	37563	37563	37563	37563
$\log \tilde{\mu}_{cp}^t$	0.101***	0.052***	0.124***	0.040***	0.040**	0.061***
	(0.018)	(0.018)	(0.017)	(0.015)	(0.015)	(0.017)
Adj. R^2	0.00	0.05	0.02	0.05	0.06	0.07
N	37570	37570	37570	37570	37570	37570
$\log \omega_{cp}^t$	0.014***	0.076***	0.009***	0.005	0.006	0.050***
	(0.002)	(0.012)	(0.002)	(0.006)	(0.006)	(0.014)
$\log \tilde{\mu}_{cp}^t$	0.042**	0.056***	0.065***	0.035**	0.034**	0.042***
	(0.016)	(0.015)	(0.019)	(0.014)	(0.013)	(0.015)
Adj. R^2	0.01	0.07	0.02	0.05	0.06	0.08
N	37563	37563	37563	37563	37563	37563
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports three sets of regressions following Equation 11: the first, including only phenotypic density ($\log \omega_{cp}^t$); the second, only genotypic proximity ($\log \tilde{\mu}_{cp}^t$); the third, both. The dependent variable follows the definition in Equation 10. Standard errors clustered at country and product level. Significance is indicated by *(10%), **(5%), and ***(1%).

Table B.6: Capability-enhanced appearance regressions (including density)

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \tilde{\mu}^t_{cp}$	0.036***	0.053***	0.047***	0.025**	0.026**	0.034***
	(0.008)	(0.009)	(0.012)	(0.012)	(0.012)	(0.013)
req^t_{cp}	-0.519***	-0.175**	-0.235**	-0.299***	-0.296***	-0.161
	(0.084)	(0.086)	(0.101)	(0.103)	(0.103)	(0.102)
edu^t_c	-0.615***		-0.358***	-0.426***	-0.351***	
	(0.097)		(0.113)	(0.116)	(0.117)	
$req^t_{cp} \times edu^t_c$	0.219***	0.084**	0.121***	0.148***	0.147***	0.082*
	(0.037)	(0.038)	(0.043)	(0.044)	(0.044)	(0.044)
$\log \omega^t_{cp}$	0.014***	0.076***	0.009***	0.005*	0.006**	0.050***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.004)
Adj. R^2	0.02	0.07	0.03	0.06	0.06	0.08
N	31920	31920	31920	31920	31920	31920
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports the regression described in Equation 12: the regressors are genotypic proximity ($\log \tilde{\mu}^t_{cp}$), the (weighted) average years of education necessary for missing occupations (req^t_{cp}, in logs), the country's average years of education (edu^t_c, also in logs), the interaction term, and density ($\log \omega^t_{cp}$). The dependent variable follows the definition in Equation 10. Robust standard errors in parentheses. Significance is indicated by *(10%), **(5%), and ***(1%).

Table B.7: Capability-enhanced appearance regressions (alternative measurement)

	(1)	(2)	(3)	(4)	(5)	(6)
$\log \tilde{\mu}^t_{cp}$	0.100***	0.051***	0.124***	0.032***	0.034***	0.056***
	(0.007)	(0.009)	(0.009)	(0.012)	(0.012)	(0.013)
req^t_{cp}	-0.049***	-0.054***	0.015	-0.003	-0.003	0.004
	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)	(0.010)
edu^t_c	0.008***		-0.000	0.010*	0.031***	
	(0.003)		(0.003)	(0.005)	(0.007)	
$req^t_{cp} \times edu^t_c$	0.021***	0.021***	0.013***	0.016***	0.015***	0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Adj. R^2	0.01	0.06	0.03	0.06	0.06	0.07
N	31835	31835	31835	31835	31835	31835
Controls		ct	pt	c,p	c,p,t	ct,pt

Notes: The table reports the regression described in Equation 12: the regressors are genotypic proximity ($\log \tilde{\mu}^t_{cp}$), the (weighted) share of missing occupations that require tertiary education (req^t_{cp}, in logs), the country's share of population with tertiary education (edu^t_c, also in logs), and the interaction term. The dependent variable follows the definition in Equation 10. Robust standard errors in parentheses. Significance is indicated by *(10%), **(5%), and ***(1%).

C Additional analysis on the role of capabilities in the development process

C.1 Periphery Trap

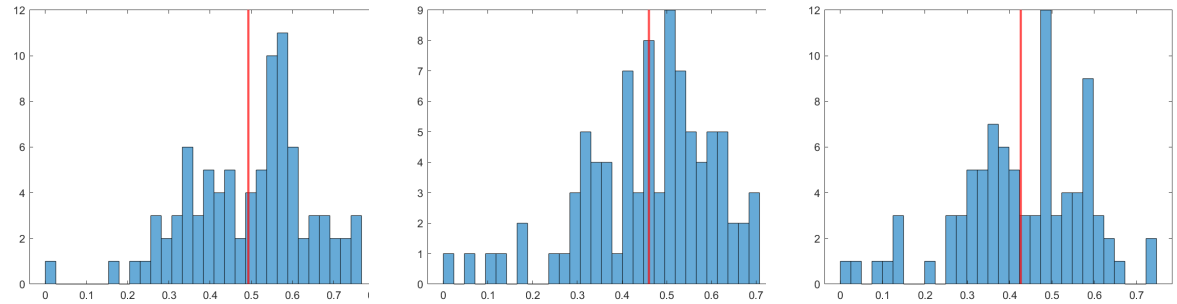
Industries differ not only in the number and composition of capabilities they require but consequently also in their position among the network of industries. In turn, this matters for a country's prospects to grow out of an industry: More central industries provide many diversification opportunities into nearby industries while such opportunities are scarce when growing out of more peripheral industries. In turn, this implies that ceteris paribus countries that are located in the periphery of the product space have worse diversification prospects. This observation is not new and it is known at least since Hidalgo et al. (2007). But our genotypic analysis allows for a novel perspective on the empirical relevance of these arguments as we now explain.

Figure C.1 shows for a selection of 6 industries the distribution of distances of a country from the remaining industries if that country was only in the respective industry. The mean distance is indicated by a red vertical line. The industries are: the industry with the fewest occupations (3161 – leather and hide tanning); two other industries in the textile cluster (3162 – footwear manufacturing; and 3159 – apparel accessories manufacturing); tobacco manufacturing (3122), which leads on the lower-right diversification path in Figure 3; fruit and vegetable preserving (3114), which is a downstream industry of agriculture and in the food cluster in Figure 3; and a mining industry (2122 – metal ore mining). All of these industries are rather peripheral in the product space but still, there are important differences in terms of how connected they are: On average, fruit and vegetable preserving (3114), which is also the industry that is most diversified in terms of its occupational inputs, is connected best, while metal ore mining (2122) has the lowest connectivity with the network of industries. What is more, this industry has very few industries at low distances, which is also the case for tobacco manufacturing (3122). In turn, this might hinder a gradual diversification were a country sequentially enters industries and moves closer to the network of industries. After all, it may be more important for a country's diversification prospects to have some stepping stones in reach which allow building up new capabilities than being initially closer on average to the network of industries.

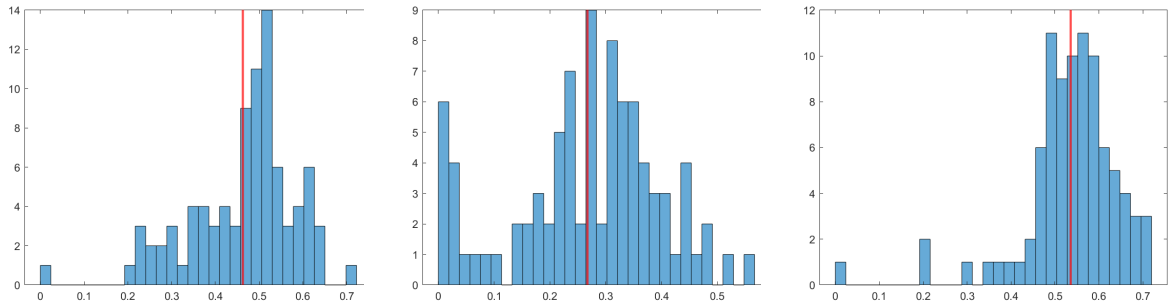
To take this point home, we consider a simple probabilistic model of diversification. Specifically, suppose that a country is able to enter a product p if its profits from doing so exceed the fixed costs of entry. Suppose further that a country's profit potential in an industry

Figure C.1: Periphery trap: Initial distances

(a) 3161 – Leather & hide (b) 3162 – Footwear manuf. (c) 3159 – Apparel accessories



(d) 3122 – Tobacco manuf. (e) 3114 – Fruit & veg. pres. (f) 2122 – Metal ore mining



Notes: The figure shows for 6 different industries as indicated in the titles of the respective panels a histogram of distances from all remaining industries for a hypothetical country that is only present in the respective industry. The vertical red line indicates the mean distance.

declines (i) in its distance from that industry and (ii) in the country's diversification. The latter reflects that countries become richer as they diversify which increases the opportunity cost for hiring labor in a new industry. The former reflects e.g. a lower productivity in an initial learning phase for new occupations. Because different occupations are complementary, this gives rise to an exponential decline of a country's productivity and, hence, profits in an industry with respect to its distance from that industry. In summary, a country c can profitably enter industry p if

$$\mu_{cp}^{*} \leq \kappa_1 \log\left(f_{cp} I_c \kappa_3 + \kappa_2\right), \tag{C.1}$$

where f_{cp} denotes the fixed cost of entry in product p as detailed momentarily, I_c the diversification of country c (i.e., the number of industries it currently has). Further details on Condition (C.1) and how it can reflect entry in richer models are provided in Appendix D.2. Here we simply note that κ_1 and κ_2 are parameters that capture 2 fundamental forces: (i) how sensitive a country's productivity in an industry is to its distance from that industry (κ_1). (ii) how difficult diversification is initially (κ_2). κ_3

governs the size of the fixed costs.

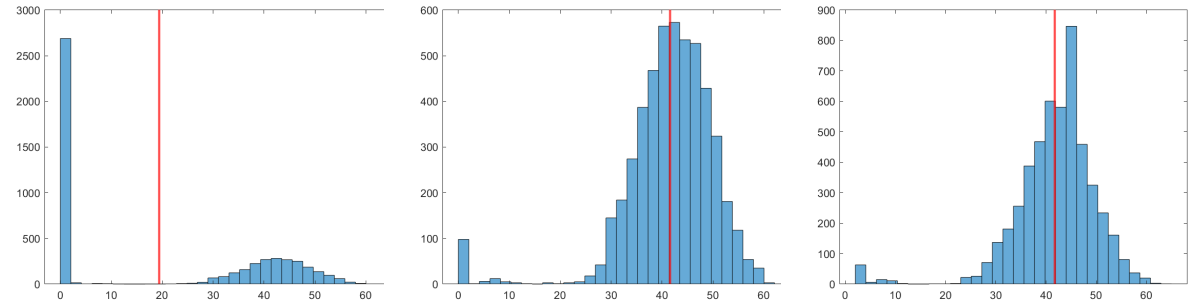
To shed light on the implications of a country's initial position in the product space for its development prospects, we use Condition (C.1). Specifically, starting in turn from each of the six industries considered in Figure C.1, we simulate 500 development paths, where at each step we draw fixed costs of entry f_{cp} from a uniform distribution on $[0, 1]$ for all industries not present in the country and then have the country enter the nearest industry among the ones satisfying Condition (C.1). If no such industry exists, further entry is not feasible, the diversification stops, and we store the number of industries the country grew into before it got stuck in this iteration. In our baseline calibration, we choose $\kappa_1 = -.1$ and $\kappa_2 = .08$, which implies that among the simulations starting from industry 3161 (leather and hide tanning), the share of countries that started to diversify and their diversification at the end of the process are broadly in line with the data.[22]

Figure C.2 plots for each of the six industries a histogram of diversifications at the end of the simulated development path across the 500 iterations. Robustness checks are provided in Appendix C.3. The figure reveals striking differences across the industries: First, countries have lower prospect to grow out of industries 3122 and 2122 when compared to industry 3161 even though these industries are much more diversified in terms of their occupational inputs and industry 3122 also is closer on average to the rest of the network—see Figure C.1. Interestingly, the poor diversification prospects when starting from 2122—see panel (f)—points to a novel type of resource curse that does not operate through higher wages—the classical argument—but through the peripheral location of metal ore mining in the product space. Second, countries are much more likely to grow out of industry 3162 than out of industry 3122 even though they have about the same average distance from the network of industries, highlighting the importance of nearby stepping stones that can lead on a pathway to prosperity. Third, 3114, which is by far closest to the network of industries and uses the most occupations nevertheless leads to similar levels of diversification when compared to industries from the textile cluster—3162 and 3159: When compared to these industries, the risk of getting stuck at very early stages of diversification is somewhat lower in industry 3114. On the contrary, when successfully diversifying, the development process tends to stop at somewhat lower levels of diversification. This is a reflection of the more peripheral structure of the natural development path out of 3114—see Figure 3. We will get back to this point when discussing entry traps next.

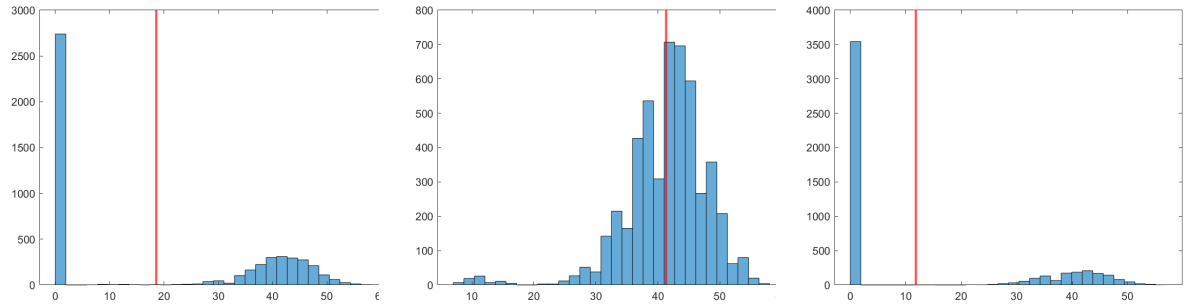
[22]Specifically, among the set of countries with less than 20 industries in 1992 to 1996 in our sample, 30% diversified their export by at least 5 industries and 50% compared to their baseline diversification. Moreover, the most diversified countries in our sample have around $40 - 50$ industries—see Figure 6(b).

Figure C.2: Periphery trap: Simulated development paths

(a) 3161 – Leather & hide (b) 3162 – Footwear manuf. (c) 3159 – Apparel accessories



(d) 3122 – Tobacco manuf. (e) 3114 – Fruit & veg. pres. (f) 2122 – Metal ore mining



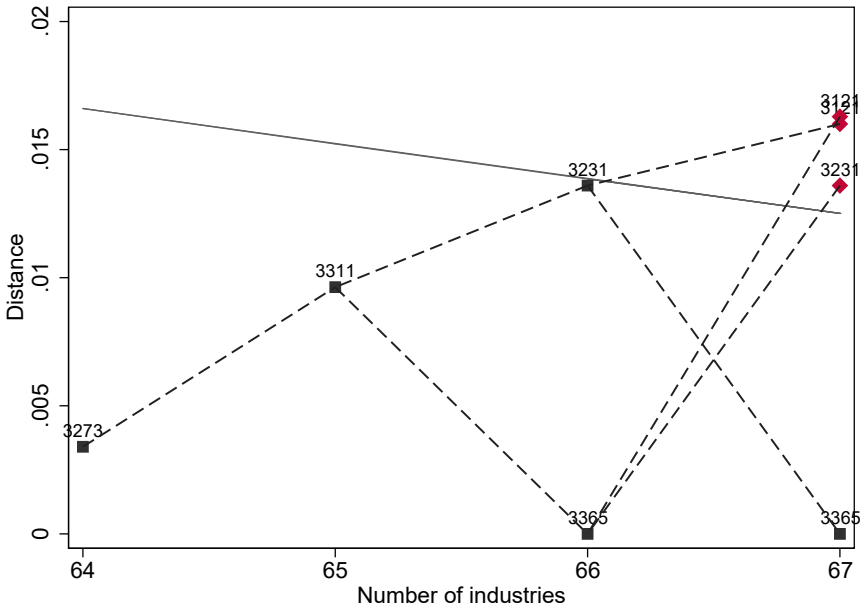
Notes: This figure summarizes the simulation as detailed in the text. Each panel shows a histogram of the number of successful diversification jumps across 10'000 simulated development paths starting from the respective industry as indicated in the title of the panel. The vertical red line indicates the mean.

C.2 Entry Trap

An issue closely related to but distinct from the periphery trap is an entry trap. The periphery trap can arise if a country is located in the periphery of the product space at early stages of development. As opposed to that, an entry trap can arise at later stages of development and it is a consequence of the two opposing effects of entry: On the one hand, economic diversification makes available additional capabilities and, hence, brings a country closer to its missing industries. On the other, diversification leads to growth and, hence, higher wages, which lowers profits from entering in new industries. An entry trap can occur if the latter effect dominates.

We can use our previous arguments to illustrate the potential importance of entry traps. Figure C.3 zooms into Figure 6(c) at the range of 64 to 67 industries. To this figure, we have added the right-hand side of Condition (C.1) using our parameter values from before, but holding constant the fixed cost f at .059 to make entry into different industries directly comparable. With 65 industries, the closest still missing industry is 3365. This is

Figure C.3: Entry trap



Notes: Simulated environment of growth by diversification. The horizontal axis represents time. On the vertical axis we represent the two sides of the inequality in **??**: the solid grey line is the right-hand side - thus representing the number of industries present in country c. The left-hand side is plotted with dots. Entry is possible only when the dot is above the solid line (highlighted with red diamonds). While in Figure C.2 we assume that the simulated country enters the closest industry, here we show other feasible paths, one of which leads to stagnation. The labels indicate the 4-digit Naics code of the added industry. The industries involved are: 3113 - Sugar and Confectionery; 3114 - Fruit and Vegetable Preserving; 3118 - Bakeries and Tortilla; 3119 - Other Food; 3122 - Tobacco.

the one considered in Figure 6(c). Given our parameter choices, there is, however, another industry 3231 that the country might grow into. Entry is often not a top-down policy decision, but the result of firms successfully organizing themselves to become competitive in something that is new to them. It is, thus, entirely possible that the firms managing to do so faster are not the ones entering the industry that is closest. Interestingly, in our illustrative example that would be beneficial for growth as there are additional diversification opportunities when entering 3231 in step 66 but not when entering 3365.

We can also use our simulation from before to shed light on potential bottlenecks in the network of industries. To that end, we summarize across all 3000 iterations—500 for each starting industry in Figure C.2—and for every industry how likely a country is to keep on diversifying upon entering an industry. Table C.1 shows the top and bottom 10 industries. Interestingly, the top 10—all of which (almost) always allowed further diversification upon entry—are manufacturing industries at or close to the core of the product space. Conversely, the bottom 10 industries are more peripheral and either rather advanced (3391 and 3364) or in the periphery of the product space (e.g. 3231, 2122, 3241)—see Figure 3.

Table C.1: Top and bottom 10 industries by prospects of further diversification

NAICS	Description	Share
3322	Cutlery & handtool manufacturing	1.000
3352	Household appliance manufacturing	1.000
3325	Hardware manufacturing	1.000
3321	Forging & stamping	1.000
3359	Other electr. equipm. & component manufacturing	1.000
3327	Screws, nuts, and bolts	0.999
3336	Engine, turbine, and power transmission equipment	0.998
3334	Ventilation, hearing, air conditioning	0.998
3361	Motor vehicle manufacturing	0.997
3332	Industrial machinery manufacturing	0.996
⋮	⋮	⋮
3365	Railroad rolling stock manufacturing	0.893
3366	Ship & boat building	0.891
3254	Pharmaceutical and medicine manufacturing	0.887
2122	Metal ore mining	0.884
3231	Printing support activities	0.872
3364	Aerospace product & parts manuf.	0.872
3262	Rubber product manufacturing	0.870
3241	Petroleum & coal products manuf.	0.865
3391	Medical equipment manufacturing	0.855
2111	Oil & gas extraction	0.781

Notes: This table shows for our previous simulation the top and bottom 5 industries in terms of the following ratio:

$$\frac{\text{\# iterations entered industry and continued to diversify}}{\text{\# iterations entered industry}}.$$

The ratio is computed over all 3000 iterations—500 for each panel in Figure C.2.

2122 and 3241 are extractive industries or their downstream industries, which further supports our previous argument that the peripheral location of these industries in our genotypic product space can give rise to a novel type of resource curse. These industries require skills that are rather specific and hence, are of little use for future diversification into new industries. Note that this is different from the conventional view on the resource curse. Loosely speaking, the resource curse is about the *quantity* of production factors being tied up in extractive industries, which lowers competitiveness in other industries. We treat all industries symmetric in this sense. By contrast, in our simulations, these industries can hinder future diversification because of the *type of skills* they require.

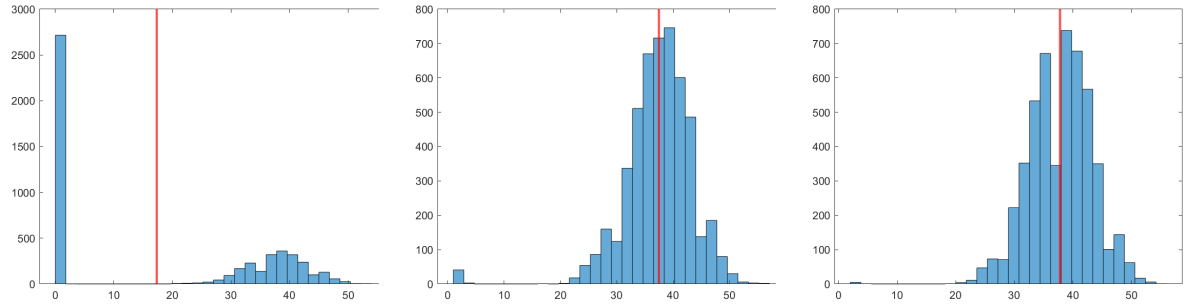
C.3 Robustness of Simulation in Section C.1

In this appendix, we present three robustness checks for Figure C.2. Figure C.4 presents simulations where—instead of jumping to the nearest feasible industry—the country jumps to a randomly selected feasible industry. Figure C.5, uses distances based on employment shares as opposed to wage-bill shares. Lastly, Figure C.6 considers a case where the fixed cost of entry increase sub-linearly with the industrial diversification, that is, where we modify Equation (C.1) as follows

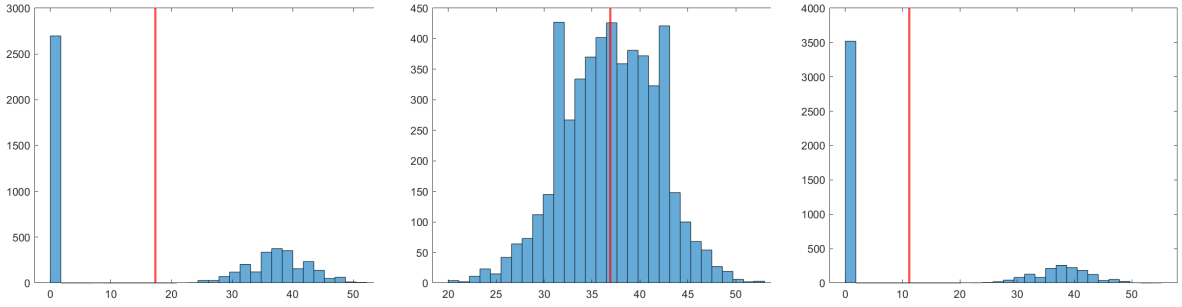
$$\mu^*_{cp} \leq \kappa_1 \log\left(f_{cp}(I_c)^{.5}\kappa_3 + \kappa_2\right). \tag{C.2}$$

Figure C.4: Periphery trap: Simulated development paths—robustness: random entry

(a) 3161 – Leather & hide (b) 3162 – Footwear manuf. (c) 3159 – Apparel accessories



(d) 3122 – Tobacco manuf. (e) 3114 – Fruit & veg. pres. (f) 2122 – Metal ore mining



Notes: This figure provides a robustness check for Figure C.2 where at each step the country enters a randomly selected feasible industry. The figures are based on 5'000 iterations each.

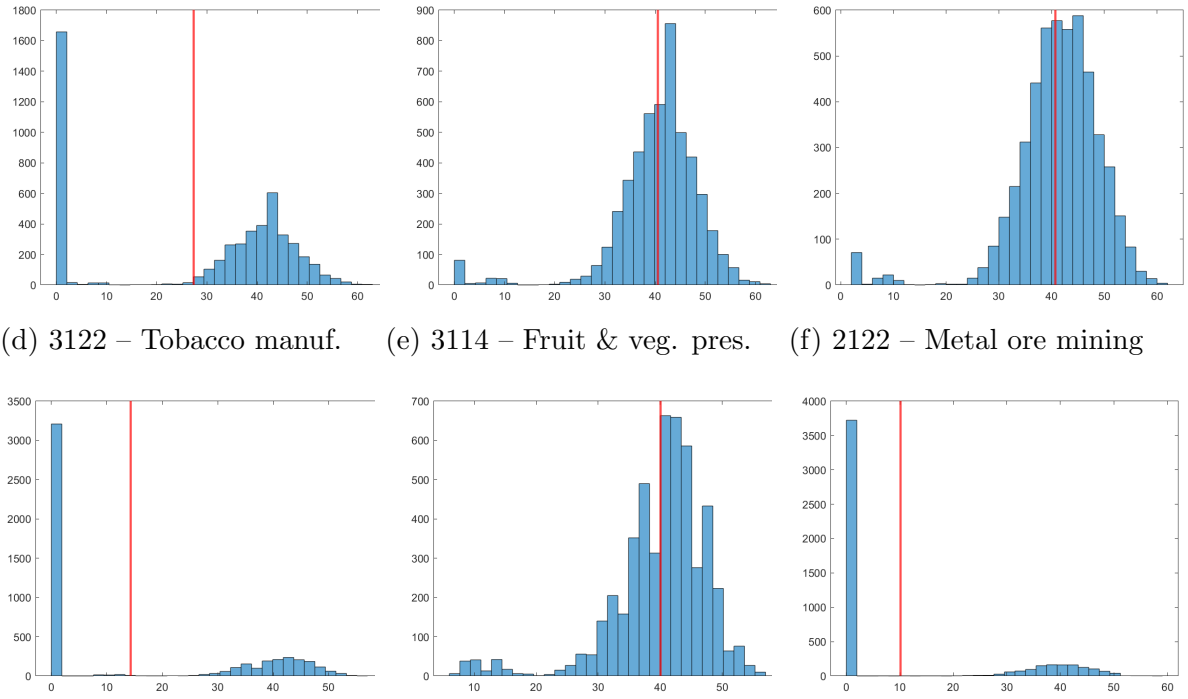
D Technical Details

D.1 Relation Between Phenotypic and Genotypic Measures

In this appendix we show (i) that the phenotypic measures entail important information about proximity and density in a capabilities-based world but that (ii) they do not in

Figure C.5: Periphery trap: Simulated development paths—robustness: employment shares

(a) 3161 – Leather & hide (b) 3162 – Footwear manuf. (c) 3159 – Apparel accessories



(d) 3122 – Tobacco manuf. (e) 3114 – Fruit & veg. pres. (f) 2122 – Metal ore mining

Notes: This figure provides a robustness check for Figure C.2 where distances have been computed based on employment shares instead of wage-bill shares. The figures are based on 5'000 iterations each.

general correctly measure technological proximities.

Considering proximities between pairs of products first, it is well known that the phenotypic proximity in Equation (3) can be re-expressed as a minimum conditional probability

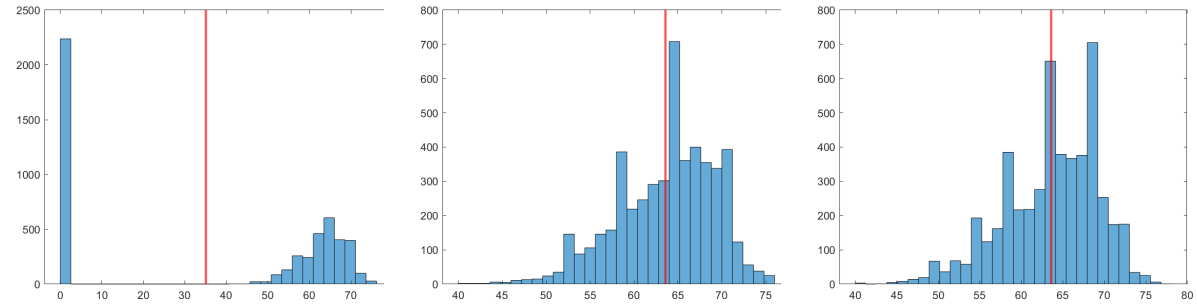
$$\begin{aligned}\Phi_{pp'} &= \frac{\sum_c M_{cp} M_{cp'}}{\max(\sum_c M_{cp}, \sum_c M_{cp'})} \\ &= \min\left\{\Pr[p'|p]; \Pr[p|p']\right\}.\end{aligned}$$

$\Pr[p'|p]$ denotes the probability that a country exports product p conditional on it exporting product p'. Interestingly, this measure does indeed entail important information on the capability overlap between pairs of products, at least if capability endowments are random. In particular, suppose that all countries successfully export all products for which they have all the required capabilities. Suppose further that a country c has any given capability with probability $r_c \in (0,1)$, analogous to Hausmann and Hidalgo (2011). Then, the unconditional probability that the country can make a product p that requires n_p capabilities is

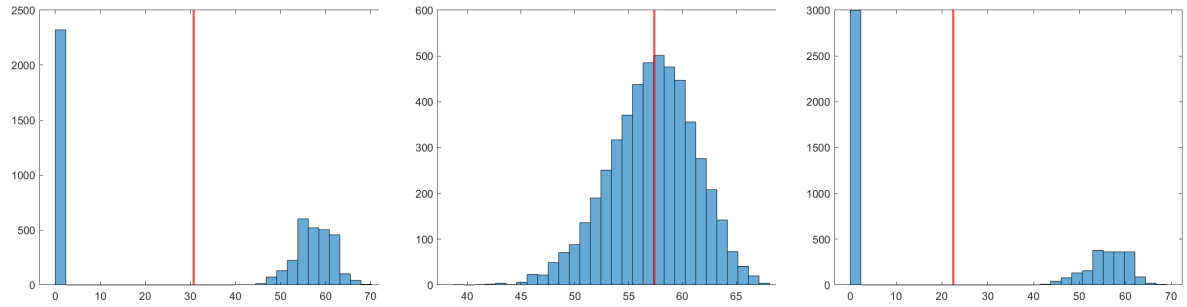
$$\Pr[M_{cp} = 1] = (r_c)^{n_p}.$$

Figure C.6: Periphery trap: Simulated development paths—robustness: Equation (C.2)

(a) 3161 – Leather & hide (b) 3162 – Footwear manuf. (c) 3159 – Apparel accessories



(d) 3122 – Tobacco manuf. (e) 3114 – Fruit & veg. pres. (f) 2122 – Metal ore mining



Notes: This figure provides a robustness check for Figure C.2 using Equation (C.2). The figures are based on 5'000 iterations each.

Conditional on making p', we know that a country has all the capabilities needed to make p'. Hence, conditional on exporting p', the probability that country c can also make product p is

$$\Pr[M_{cp} = 1 | M_{cp'} = 1] = (r_c)^{n_{p \neg p'}} ,$$

where $n_{p \neg p'}$ denotes the number of capabilities that p requires but p' not. The key point to note is that this probability is increasing in the capability overlap between p and p'. This means that if all products require the same number of capabilities—and, thus, $\Pr[p'|p] = \Pr[p|p']$—the phenotypic proximity is increasing in the genotypic proximity. More generally, this suggests that the phenotypic proximity entails important information about the capability overlap between pairs of products, albeit it typically does not measure this overlap correctly.

Additional problems arise when it comes to density as this requires aggregating pairwise distances between products into a distance between a country and a product, i.e., a basket of products and a single product. In general, this requires accounting for the fact that the products in a country's export basket differ in their respective capability overlaps.

The genotypic approach does so by backing out country capability endowments first. The phenotypic approach instead does not so in general, unless strong restrictions are imposed.[23]

D.2 Further details on Condition (C.1)

In this appendix, we show how Condition (C.1) captures in a simple way the key entry dynamics in Diodato et al. (2022).

Diodato et al. (2022) consider a Small Open Economy (SOE) that grows by diversifying its export basket, such that the wage rate is proportional to ${I_S}^{1/\sigma}$, where I_S is the number of industries in the SOE and σ a parameter capturing the gains from industrial diversification.[24] Entering a new industry involves a fixed cost in terms of labor, i.e., this fixed cost is proportional to the wage rate. Entry further requires training workers in all occupations that are new to the SOE, and the productivity of these workers is lowered by a factor $\lambda < 1$ in an initial learning period. Upon entry, a firm makes profits that are thus decreasing in its distance μ^*_{Sp} from that industry and, quite intuitively, in the wage rate. Taken together, firms in the SOE find it profitable to enter industry p if

$$\gamma \left[\lambda^{\mu^*_{Sp}}\right]^{\sigma-1} \frac{1}{I_S} \geq f_{Sp},$$

where γ is a constant, $\sigma - 1$ governs how sensitive profits are to the SOE's distance from industry p, and f_{Sp} are the fixed costs of entry.[25] Taking logs and re-arranging terms, we get

$$\mu^*_{Sp} \leq \underbrace{\frac{1}{(\sigma-1)\log(\lambda)}}_{\kappa_1} \kappa_1 \log\left(f_{Sp} I_S \underbrace{\gamma^{-1}}_{\kappa_3}\right),$$

where the inequality gets reversed because $\log(\lambda) < 0$. Condition (C.1) generalizes this by introducing an additional additive term for the fixed cost, $\frac{\kappa_2}{I_S}$ such that total fixed cost of entry are

$$f_{Sp} + \frac{\kappa_2}{I_S}.$$

[23]The easiest way of seeing this is by considering two products A and B with the same capability requirements. Then, phenotypic approaches will generally ascribe a country greater density for a third product C if it has both products A and B compared to when it has only one of them, while conditional on having A adding B does not add to a country's capability endowments and vice versa.

[24]They consider an Armington (1969)-type model where each country is equipped with a distinct variety in each industry and where these varieties are then aggregated in a CES consumption aggregator to industry bundles. In such case σ is the constant elasticity of substitution in consumption.

[25]This condition exactly maps onto Condition (B.2) in Diodato et al. (2022) if the fixed cost of entry and the learning cost have a small effect on the equilibrium wage in the SOE in the entry period—see their Equation (13).

This term governs how hard the initial jump on a pathway to prosperity is.