INTERNATIONAL MONETARY FUND

High Performance Export Portfolio: Design Growth-Enhancing Export Structure with Machine Learning

Natasha Che, Xuege Zhang

WP/22/75

2022
APR

WORKING PAPER

 WP/22/75

IMF Working Paper
Asia & Pacific Department

High Performance Export Portfolio: Design Growth-Enhancing Export Structure with Machine Learning

Prepared by Natasha Che, Xuege Zhang*

Authorized for distribution by Shanaka Jayanath Peiris
April 2022

ABSTRACT: This paper studies the relationship between export structure and growth performance. We design an export recommendation system using a collaborative filtering algorithm based on countries' revealed comparative advantages. The system is used to produce export portfolio recommendations covering over 190 economies and over 30 years. We find that economies with their export structure more aligned with the recommended export structure achieve better growth performance, in terms of both higher GDP growth rate and lower growth volatility. These findings demonstrate that export structure matters for obtaining high and stable growth. Our recommendation system can serve as a practical tool for policymakers seeking actionable insights on their countries' export potential and diversification strategies that may be complex and hard to quantify.

JEL Classification Numbers:	F1, F4, O1, O4
Keywords:	export diversification, comparative advantage, machine learning, collaborative filtering, economic growth, international trade
Author's E-Mail Address:	nche@imf.org, xuegez@andrew.cmu.edu

* The authors would like to thank Davide Furceri and participants in the APD and WHD departmental seminars for their helpful comments. The views are entirely our own.

High Performance Export Portfolio: Design Growth-Enhancing Export Structure with Machine Learning

Natasha Che[*], Xuege Zhang[†]

Abstract

This paper studies the relationship between export structure and growth performance. We design an export recommendation system using a collaborative filtering algorithm based on countries' revealed comparative advantages. The system is used to produce export portfolio recommendations covering over 190 economies and over 30 years. We find that economies with their export structure more aligned with the recommended export structure achieve better growth performance, in terms of both higher GDP growth rate and lower growth volatility. These findings demonstrate that export structure matters for obtaining high and stable growth. Our recommendation system can serve as a practical tool for policymakers seeking actionable insights on their countries' export potential and diversification strategies that may be complex and hard to quantify.

Keywords: export diversification, comparative advantage, machine learning, collaborative filtering, economic growth, international trade

JEL Codes: F1, F4, O1, O4

[*]International Monetary Fund. Email: nche@imf.org
[†]Tepper School of Business, Carnegie Mellon University. Email: xuegez@andrew.cmu.edu

1 Introduction

Over the past decades, many success stories of growth and income convergence- most notably, China and several other East Asian emerging economies- have been export-led. Some of these countries have governments that actively pursue industrial policies that foster strategic export industries; others let the market take the lead. Setting aside the pros and cons of each approach, there is no denying that export diversification and industrial structural change are important for growth (e.g., Aiginger and Rodrik, 2020). Despite the topic's relevance, however, there is surprisingly little practical guidance from the economic literature regarding what types of export structures are suitable for growth, and what specific products each country should consider diversifying into.

Classical trade theories suggest that countries should export what they are relatively good at producing, i.e., following comparative advantages. But how exactly does one ascertain comparative advantages? Trade theories predict that many developing countries tend to have comparative advantages in labor-intensive exports and should, to some extent, stay away from capital-intensive industries. But in reality, comparative advantages contain far more dimensions than capital and labor. Some of these dimensions are quantifiable in a more linear way with production technologies, whereas others are not.

The matter becomes even more complicated when we consider the fact that comparative advantages evolve as a country grows. How could the export structure change as a result to achieve better growth performance? General trade theories and empirical studies do not go very far in providing country-specific diversification strategy, or practical insights in guiding the structural change in exports.

In a recent study, Che (2020) proposes a novel method to operationalize the concept of comparative advantage and its evolution. It uses collaborative filtering algorithms in machine learning most commonly applied to product recommendations in e-commerce, to produce export diversification recommendations that reflect a country's *latent comparative advantages* and future potentials in export structure. Section 3 will go over the details of the methodology. But the

basic intuition is that a country is likely to have comparative advantages in those products that are highly related to the products it is already good at exporting (i.e., products with *revealed comparative advantages*), where the "relatedness" between any two products is measured by the similarities among countries that are the main exporters of the two products. Moreover, it turns out that the export structures recommended by the "product-based KNN" algorithm can predict the evolution of actual export structure for several high-growth countries (including China, India, Chile and Poland) reasonably well. Here the *export structure* is measured by the number of products recommended by the algorithm, categorized by Standard International Trade Classification (SITC) 4-digit codes, that belong to each of the 10 SITC 1-digit sectors, as a share of the total number of recommended products.

Inheriting the intuition of exporting based on comparative advantages, this paper further provides cross-country evidence by designing a recommendation system that can serve as a practical tool for policymakers to seek actionable insights for diversification strategies. The rationale for such an export recommendation approach comes from two cross-country observations. First, products that require similar production inputs and know-how tend to show up in an export portfolio together. For example, a country that has successfully exported beef can branch into, with some effort, dairy. A country that has mastered the trade of exporting desktop computer hardware is in a better position to produce and export cellphones, than otherwise. Therefore, the products in a country's existing export portfolio contains valuable information regarding what other products the country can get good at producing. Second, countries with similar comparative advantages tend to export similar products. Bangladesh and Vietnam are both successful in exporting garments because of the countries' shared abundance in low-cost labor. New Zealand and Uruguay both specialize in cattle exports partly because of the high availability of pasture land. In other words, products related to a country's existing exports and export portfolios of *similar countries* offer useful information about the country's latent comparative advantages, even though the latter cannot always be neatly expressed quantitatively.

It is necessary to note that our key index calculation (i.e., the *similarity score*, which will

be specified later) is based on *export portfolio*, instead of *export diversification* as commonly understood. This is because the export portfolio is more general to address the evolution of trade patterns in a cross-country analysis. In addition, it can also provide insights on diversification strategies as an application, especially in a single-country analysis. A country can double the number of products it exports, i.e., diversification in numbers, without changing its export structure at all, if the sectoral distribution of its exports stays the same. In contrast, if a country used to export 100 products all in the food industry, but now switches to exporting 50 products in the food sector and 50 in the chemical industry, it has not "diversified" in numbers, but its export structure has changed. A country's export portfolio can be potentially improved by diversification in the number of export products, as well as adjusting the export structure to fit its evolving comparative advantages. Moreover, the expression "diversification" per se may not be well-defined when comparing different countries. For example, it can be controversial whether a country already producing a fair amount of products becomes relatively more diversified when its exporting products expand within a sector, compared to other countries with the same change in the total number of products. Therefore, in our cross-country analyses, we use the *export portfolio* to capture more general features on the evolution of trade. Note that we do not take a stand ex-ante regarding whether a specific country should consider a structural change or a diversification. The machine-learning-based export recommendations can provide useful guidance on both.

In this paper, our primary goal is to test the hypothesis that export product recommendations based on a collaborative filtering algorithm indeed reflect what a given country's export structure could look like to help it grow better at any given time. To do this, we first use a product-based k-nearest neighbors algorithm similar to Che (2020) to make annual export product recommendations in the SITC 4-digit product space, for over 190 economies over three decades. We then compare the recommended sectoral structure of exports with the actual export structure of each country. See Section 3 for details on methodology.

If the export recommendations produced by the algorithm indeed capture countries' latent

comparative advantages, we should observe that countries whose export structure closely aligns with the recommended structure would have better growth performance. Here we define "better" as higher growth and lower growth volatility.

A preliminary look at the data appears to support our hypothesis. Figure 1 plots the cross-country correlation between average real GDP growth per capita over 1982-2017 and average *similarity score* between a country's actual export structure and recommended export structure produced by our algorithm.[1] Figure 2 plots the correlation between the 5-year standard deviation of annual growth rate and the similarity score. The charts indicate that countries with an export structure closer to the recommended structure enjoy higher growth, lower growth volatility and higher risk-adjusted growth. The same pattern can be discerned from Figure 3, which shows a positive correlation between the similarity score and countries' risk-adjusted growth, i.e., 5-year average growth divided by standard deviation of growth.

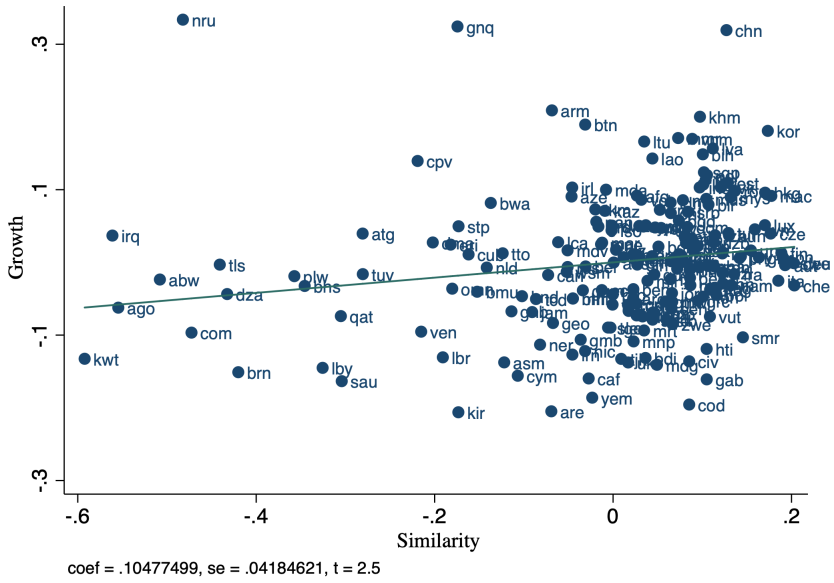


Figure 1: Relationship between "similarity score" and growth

[1]The similarity score is calculated as the Pearson correlation between actual and recommended export structures. Thus it has a theoretical range of [-1, 1]. See Section 3 for details.

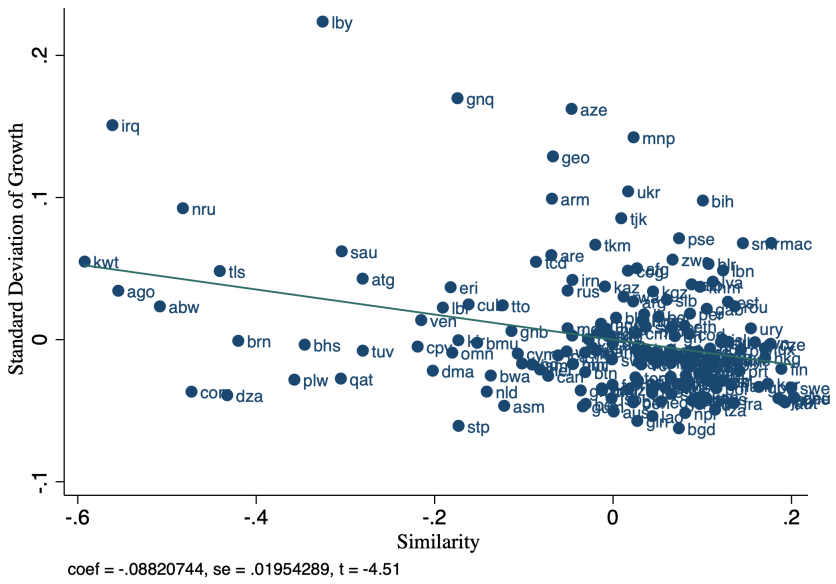


Figure 2: Relationship between "similarity score" and growth volatility

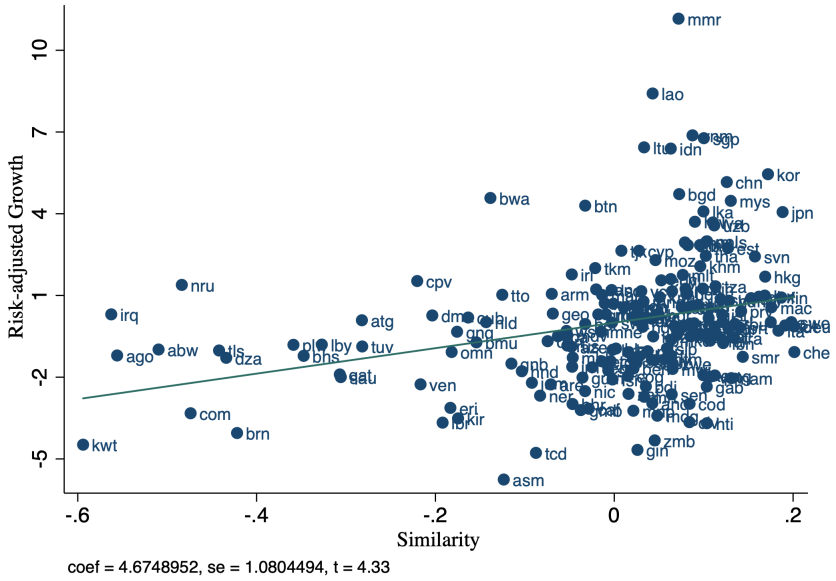


Figure 3: Relationship between "similarity score" and volatility-adjusted growth

It is interesting to look at some examples of specific countries as well. Figure 4 plots the evolution of the similarity score between algorithm-recommended export structure and actual export structure for China, Singapore, South Korea, and Germany.[2] Since late 1980s, the similarity

[2]The similarity scores are calculated annually and the charts present 5-year moving averages of the scores.

score for China has increased significantly, from below the world average to top 3% of the world sample. The magnitude of increase for Singapore is similar. For South Korea, though the similarity score has a decreasing trend in general (although increased again in the most recent years), it still remains at high levels (top 10%) on average. Likewise, Germany has one of the highest similarity scores in the world, which is unsurprising given the country's diversified and dynamic industrial export base over the past decades.

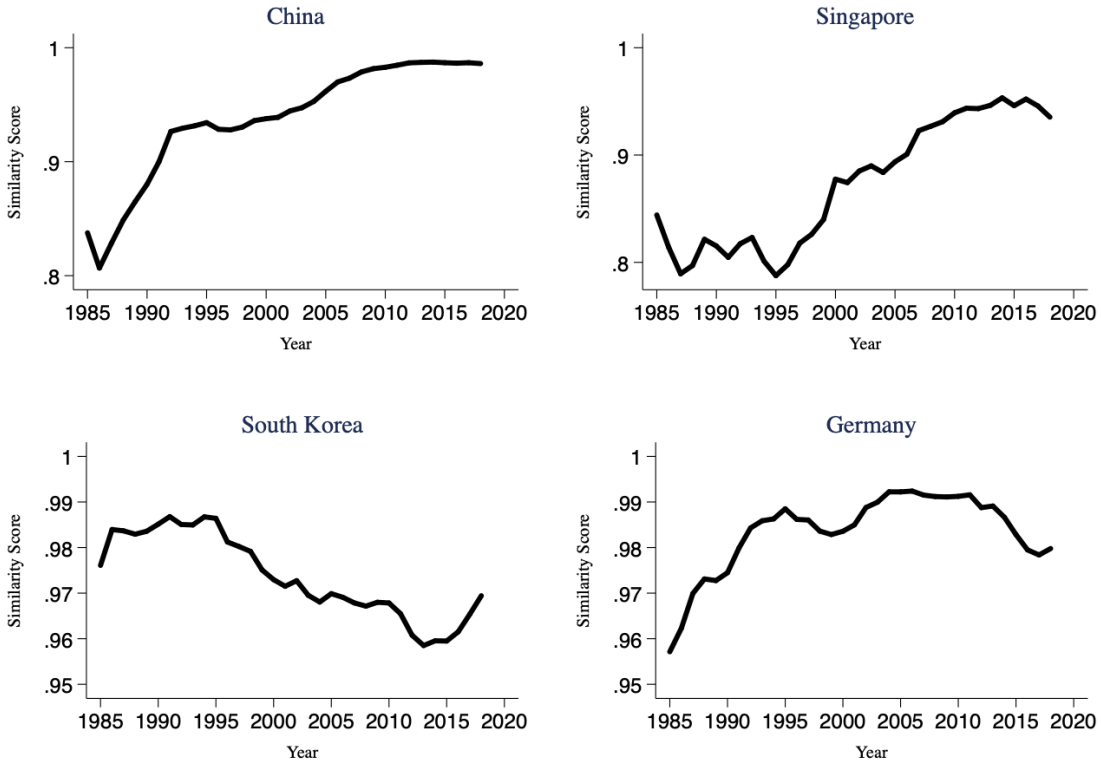


Figure 4: Similarity scores for select high-growth & developed countries (5 year MA)

Figure 5 plots the evolution of the similarity score for several developing countries with lower growth– Honduras, Kuwait, Libya, and Venezuela. For Libya and Kuwait, the similarity scores are particularly low. Though the score for Kuwait has increased in the past since 2010, the score remained below 0.2 until early 2010s and is still below 0.4 in recent years, compared to the world average of 0.82. For Honduras and Venezuela, the similarity scores are higher, but are still below the world level and have dropped significantly since the mid-1990s, likely reflecting a decline in

diversification and manufacturing capacity.

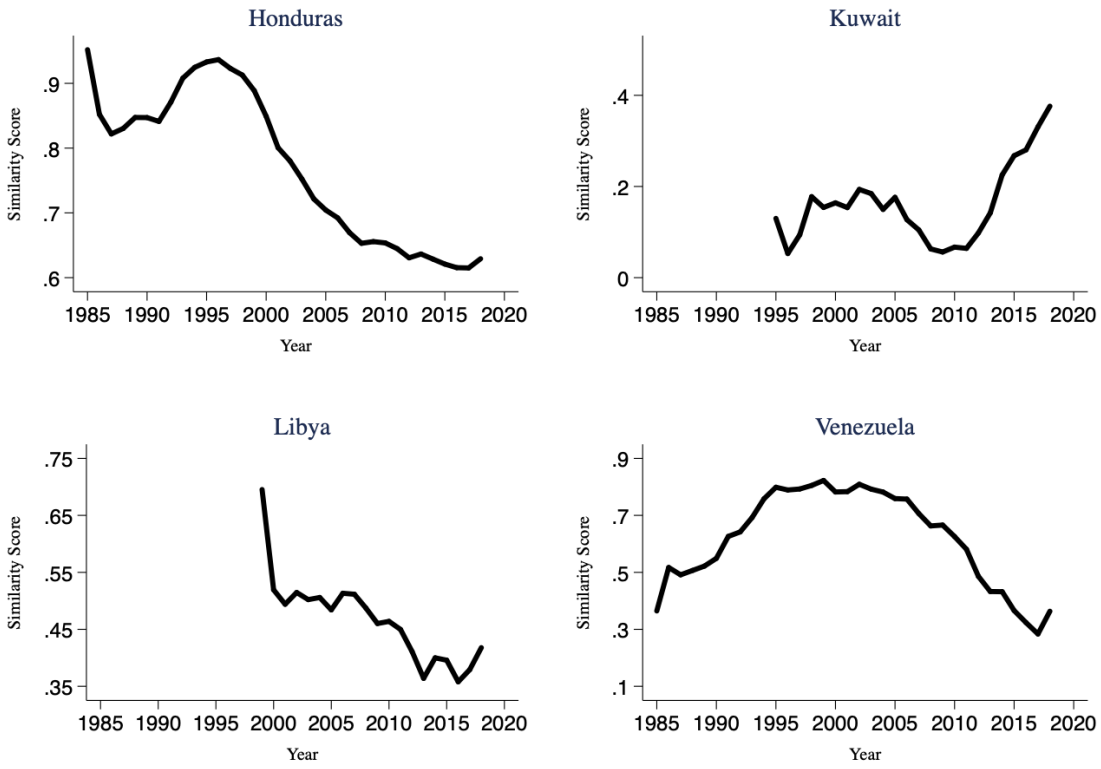


Figure 5: Similarity scores for select low-growth & fragile states (5 year MA)

The rest of the paper is organized as follows. Section 2 presents related literature on export structure and diversification. Section 3 explains the product-based KNN algorithm and our empirical methodology. Section 4 describes the data. Sections 5 and 6 present the baseline empirical results and robustness checks. Section 7 concludes.

2 Literature

The literature closest to our paper is the studies on the so-called *product space* and its implication for diversification and growth (e.g., Hausmann & Klinger 2007, Hidalgo & Hausmann 2009). Like the current paper, this strand of research seeks to understand a country's export structure by looking at the *relatedness* among products.

But there are two key differences. The first is regarding the efficiency in the use of information contained in the trade data. The product-space literature uses a probability formula to represent the relatedness, or *proximity* between two products.[3] While this approach features a clean, easy-to-understand formula and makes subsequent analyses computationally simpler, it is at the cost of not fully exploiting the information contained in the data matrix of country-product exports. In contrast, the product-based KNN algorithm in the present paper makes more efficient use of the data to detect the unique blend of characteristics of countries and products. This leads to potentially better recommendations. To be sure, it is at the cost of requiring more computational resources and forsaking the easily comprehensible linear formula. This is a common drawback of many machine learning algorithms– the nonparametric nature of the approach can make some results seem "magical" and harder to explain with linear logic.

The second, and more important, difference is one of perspectives. The product-space literature makes specific value judgments about the worthiness of different products for diversification purpose. A product's diversification value is seen as broadly depended on 1) how "complex" it is, meaning, how much sophisticated knowledge is required to product the product, and 2) how closely related the product is to other more complex products. Each product is assigned a complexity level as such. The rationale for doing so is a reasonable one– more complex products have higher value-added, use more human capital, and face less global competition. And products that are "bridges" to the more complex products may be a pathway for a country to move up the international value chain. Some empirical evidence shows that diversifying into these products is supposed to be better for growth (Hausmann, 2007). However, several issues emerge when this model is used for recommending export products to specific countries. First, there is an underlining tension between this line of thinking and the framework of comparative advantages that the product-space analysis is built on. By assigning each product a score of virtue (e.g., industrial products are good, agricultural commodities are bad), it leads to a tendency to recommend prod-

[3]Specifically, the proximity between product A and product B is defined as the probability that a country exports product A given that it exports product B, or vice versa. For example, suppose that 17 countries export wine, 24 export grapes and 11 export both, all with revealed comparative advantage. Then, the proximity between wine and grapes is 11/24 = 0.46.

ucts that the model deems universally worthy to countries of drastically different fundamentals. In the extreme, though improbable, scenario where all countries internalize the same worthiness ranking of products in developing their export structure, there would be no comparative advantages to speak of. Secondly, to come up with a tractable, universally applicable scoring system for "product complexity", strong assumptions need to be made that reduce the feature dimensions of reality and throw away valuable country- and product- specific information, which may limit the model's usefulness in producing realistic export recommendations for individual countries.

In contrast, the approach of the current paper is agnostic about the diversification value of any specific product. Instead, we seek to fully exploit the information contained in the country-product space, and make realistic export recommendations off of a country's current revealed comparative advantages. One implication is that countries do not necessarily need to chase the "complex" exports to achieve better growth performance. As Section 5 shows, countries whose export structure closely aligns with the algorithm-recommended structure have higher and more stable growth, even though the algorithm's recommendations do not make any specific judgment regarding product complexity, and are solely based on information from a country's currently revealed comparative advantages.

The paper is also related to the literature on the relationship between export diversification and countries' economic performance. Existing research asserts that export diversification is a key element in the economic development process, particularly for developing and emerging market economies trying to catch up with their advanced peers. Various studies provide evidence of a positive association between export diversification and economic development (e.g., Imbs and Wacziarg, 2003; Klinger and Lederman 2004 and 2011; Cadot et al., 2011). Numerous country studies also supports the benefits of export diversification. For example, Feenstra and Kee (2008) use data from a large set of countries exporting to the US, to show that a sustained increase in export diversification results in increases in productivity and a notable increase in the GDP of the exporters. IMF (2014) finds that diversification in exports and in domestic production has been conducive to faster economic growth in LICs. Al-Marhubi (2000) provides similar findings within

a set of developing economies. Balaguer and Cantavella-Jorda (2004) find that export variety plays a key role in Spain's economic development. And Herzer and Danzinger (2006) report a positive impact of export diversification on economic growth of Chile. Research also points to a positive association between export diversification and macroeconomic stability (e.g., IMF, 2014).

However, not all types of diversification are created equal, and diversification for its own sake is hardly a recipe for sustainable growth. A fundamental idea of the classical international trade theory is that under free trade, countries will tend to export what they are relatively good at producing, i.e., products they have a comparative advantage in. "Diversifying" into industries that are misaligned with a country's current endowment fundamentals, as the former Soviet-block nations did after World War II through industrial policies that aimed to accelerate industrialization, has negative growth consequences (Lin, 2009). On the other end of the spectrum, delayed industrialization also leads to negative growth outcomes, as the experience of many resource-rich countries that are entrenched in their over-dependence on commodity exports has shown (e.g., Frankel, 2010).

A difference in focus between the current paper and the export diversification literature is that the latter sees diversification as mostly in increasing the number of export products, while the current paper emphasizes more on the structure of the export portfolio. Our algorithm does provide a list of recommended products for each country, which offers useful insights for countries looking to increase the number of export items. But our econometric exercise focuses on the growth impact of the appropriate export structure, i.e., sectoral distribution of exports.

3 Methodology

The goal of this paper is to answer the question of whether our algorithm-based export recommendations can result in growth-enhancing export structures, in the sense that countries that follow the recommendations could achieve better growth performance. We go about answering this question in the following steps:

- **STEP 1.** Determine the number of SITC 4-digit products to recommend for each country (see Section 3.1). The number of recommendations is derived from a country's size and development level.

- **STEP 2.** Generate a list of recommended export products for each country-year in the sample using a product-based KNN algorithm (see Section 3.2).

- **STEP 3.** Calculate the *similarity score* between the export structure implied by the list of recommended export products and the actual export structure, for each country-year (see Section 3.3).

- **STEP 4.** Estimate the impact of the similarity score on growth, volatility, and risk-adjusted growth (see Section 3.4).

An important concept used throughout the paper is *Revealed Comparative Advantage* (RCA). The RCA score, first introduced by Balassa (1965), is a popular measure to calculate the relative importance of a product in a country's export basket. Formally, the RCA score of country i in product j can be calculated as:

$$RCA_{ij} = \frac{E_{ij} / \sum_j E_{ij}}{\sum_i E_{ij} / \sum_i \sum_j E_{ij}}$$

where E_{ij} is the export value of product j from country i. $\sum_j E_{ij}$ is the total export value from country i. $\sum_i E_{ij}$ is the total export value of product j from all countries around the world. And $\sum_i \sum_j E_{ij}$ is the total world exports.

A *high-RCA product* for country i is defined as a product with its $RCA_{ij} > 1$. Mathematically, it means that the product's share in the country's export portfolio is greater than its share in the total world exports, which can be seen as an indication that the country has a comparative advantage in the product. For example, vehicle exports were about 12 percent of total world exports in 2017, while they constituted 22 percent of total exports from Mexico. Therefore, $RCA_{ij} = 22/12 = 1.8$ for Mexico's vehicle exports in 2017. Since it is greater than 1,

according to our criteria, Mexico has a revealed comparative advantage in automobiles. Or to put it another way, automobiles is a high-RCA product for Mexico. The recommendation algorithm that will be introduced in Section 3.2 essentially simulates a hypothetical RCA score for each country-product combination, and pick the top products with the highest hypothetical scores as the recommended export portfolio for country i.

3.1 Choosing the number of recommended products

Examining the export data by SITC 4-digit industry[4] reveals the following empirical regularities. *First,* more developed economies tend to have a larger number of high RCA products. This reflects the fact that more advanced economies have a wider range of production knowledge and more sophisticated production structures. Figure 6 regresses the number of high RCA exports of each country on its real GDP per capita relative to the U.S. level, controlling for the country size, and shows a positive relationship.[5] *Second,* bigger countries tend to have a larger number of high RCA exports. This is unsurprising, as population size correlates highly with the number of firms, the amount of human capital and the amount of other production resources a country may have, enabling the country to viably export a wider range of products. In addition, some industries and products need a minimum scale to be sustainable. Figure 7 plots this positive relationship between number of high RCA exports and country population.

There are obviously other factors that determine how many high RCA exports a country has. But since we are focusing on exploring export structure instead of simply expanding the number of export products, we pin down the number of recommended export products for each country as predicted by the country's size and development stage. Specifically, we run the following estimation:

$$K_{rca,it} = \beta_1 GDP_{it} + \beta_2 POP_{it} + \gamma_t + \epsilon_{it} \tag{1}$$

[4]See Section 4 for a more detailed description of the underlining data.
[5]This chart reproduces Figure 3 of Che (2020).

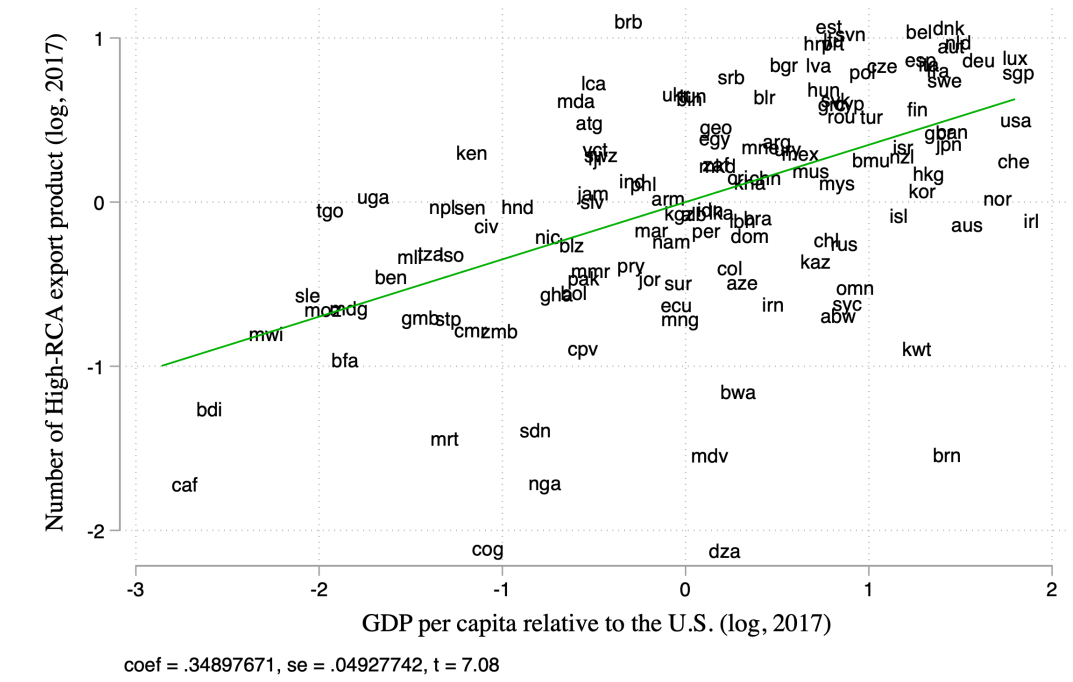


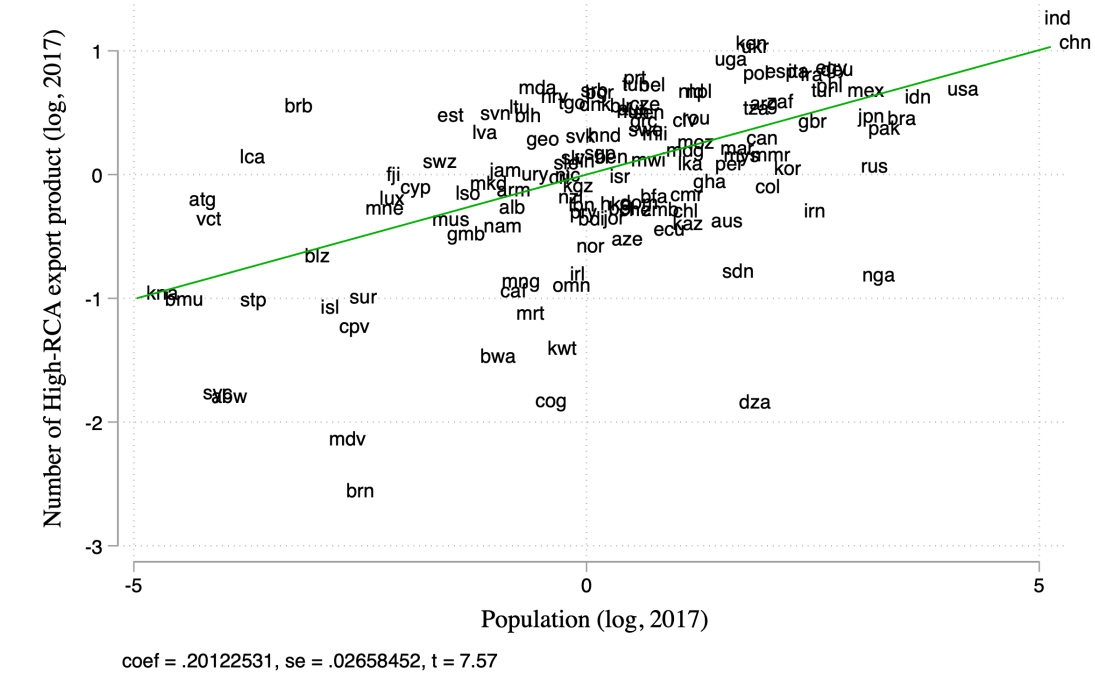
Figure 6: Number of High-RCA Exports v.s. Income Level, Partial Regression Plot

where $K_{rca,it}$ is the number of high RCA exports that country *i* has in year *t*. GDP_{it} is GDP per capita and POP_{it} is population size. We add a time fixed effect γ_t in the regression, as the average number of high RCA exports tends to rise overtime around the world, with the increase in product variety brought about by economic growth and technological change. We use $\hat{K}_{rca,it}$, the predicted number of high RCA exports from the regression, as the number of products we will recommend to each country.

3.2 The recommendation algorithm

Our export product recommendation system employs a product-based K-nearest neighbor (KNN) algorithm that is widely used in the collaborative filtering recommendation systems of online commerce.[6] The goal of the exercise is to generate, for each country-year, a list of "top-K recommendations", i.e., K products that a country should export the most of based on empirical

[6]See Che (2020) for more detailed explanations of other similar algorithms.



coef = .20122531, se = .02658452, t = 7.57

Figure 7: Number of High-RCA Exports v.s. Population, Partial Regression Plot

regularities.[7]

The algorithm produces the list by computing a *recommendation score* for each product for the underling country, using the training dataset of RCA scores by country and SITC 4-digit product, and recommending the K products with the highest recommendation scores. Here, we set K to be equal to the $\hat{K}_{rca,it}$ estimated in Section 3.1 for each country i in year t.

The underlining data used in the recommendation algorithm can be represented as a $m \times n$ matrix R, where m is the number of country-year combinations in the database, and n is the total number of SITC 4-digit products. Each element of R, i.e., r_{ij}, is country i's RCA score in product j. R is a sparse matrix due to the fact that each country only exports a subset of the products in the SITC universe.[8] In the case that country i does not export any product j, $r_{ij} = 0$. If in running

[7] In our implementation, we mainly include the fundamentals in scale and development stage for the cross-country analysis. Note that these are not (necessarily) optimal levels to implement at country level. Rather, it provides a way to systematically compare portfolio across countries based on one explanatory metrics, supported by empirical regularities. There can be alternative ways to obtain number of products a country could (or even should) export, which are more suitable and relevant when considering specific country cases.

[8] The dimension of the space depends on the granularity of products.

the algorithm, multiple years of export data are used as the training set, then each country-year is a row in R, i.e., $m = c \times y$, where c is the number of countries in the dataset, and y is the number of years included. In practice, we set $y = 1$. In other words, when we generate export recommendations for country i in 2017, only the cross-country export data for 2017 is included in the training set.[9]

KNN is one of the most frequently used methods in solving classification and pattern recognition problems, and is a popular approach in constructing recommender systems. The basic idea of KNN is learning by analogy– classifying the test sample by comparing it to the set of training samples most similar to it. Different KNN implementations vary in terms of their choices of how the similarity between input vectors is calculated. In the present paper, the cosine similarity score is used as the similarity measure.

The intuition behind the product-based KNN implementation is simple– first look at what products a country already has a revealed comparative advantage in, and then recommend other products that are "related" to those products. To explain the approach in more details, first write the RCA score matrix R as:

$$R = \begin{bmatrix} \mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n \end{bmatrix}$$

where $\mathbf{p}_j$, an arbitrary element in R, is a vector of length m that represents the RCA scores of product j for all the m countries[10] in the sample:

$$\mathbf{p}_j = \begin{bmatrix} r_{1j} \\ r_{2j} \\ . \\ . \\ r_{mj} \end{bmatrix}$$

[9] We experimented with including multiple years of data in the training set, but found no significant improvement in the results, while the model took longer to compute as the size of m increases.

[10] Note that in our implementation, m is effectively the cross-sectional country numbers. In machine learning terminology, each product in the sample has m features.

The cosine similarity between products j and j' is equal to $\frac{\mathbf{p}_j \cdot \mathbf{p}_{j'}}{\|\mathbf{p}_j\| \|\mathbf{p}_{j'}\|}$, which ranges from -1, when the two vectors are the exact opposite, to 1, when the two are exactly the same. The intuition behind this is that by comparing the two sets of countries that export i and j, and how important the products are in the countries' export baskets, information can be inferred regarding how closely related the two products are.

The implementation of the product-based KNN recommender for country i in year t involves the following steps:

1. Represent each product in the SITC 4-digit product space as a vector of RCA scores, $\mathbf{p}_j$.

2. Select the set of products that country i has a revealed comparative advantage, i.e., $r_{ij} > 1$, which will be referred as the high-RCA product set of country i.

3. For each $j \in [1, n]$, calculate the predicted value of r_{ij} as a weighted average RCA score of the high-RCA product set, weighted by the cosine similarity between product j and the products in the country's high-RCA set.

4. The recommended products for country i are the $\hat{K}_{rca,it}$ products with the highest predicted r_{ij} values (i.e., recommendation scores), where $\hat{K}_{rca,it}$ comes from the estimation in Section 3.1.

We repeat the above steps for each country-year pair to generate the recommended export portfolio in terms of SITC 4-digit products for every country in each sample year.

3.3 Calculating similarity scores

For the next step, we compute the similarity between the actual export portfolio of a country and the recommended export portfolio.

We define the portfolio structure of country *i*'s actual exports in time *t* as the number of high RCA exports (at SITC 4-digit level) that belong to each SITC 1-digit sector,[11] as a share of total

[11] See appendix Table 17 for the full list of SITC 1-digit sectors.

number of high RCA exports. In other words, let $N_{l,it}^{actual}$ be the number of high RCA exports in sector l, and

$$s_{l,it}^{actual} \equiv \frac{N_{l,it}^{actual}}{\sum_{l'} N_{l',it}^{actual}}$$

is the share of the number of high RCA exports that belong to sector l in the total number of high RCA exports. Country i's export structure, S_{it}^{actual}, is thus defined as a $L \times 1$ vector: $\left[s_{l,it}^{actual}\right]_{L\times 1}$, where $L = 10$ with the SITC sectoral level aggregation.

Similarly, we define the recommended export structure $S_{it}^{rec} \equiv \left[s_{l,it}^{rec}\right]_{L\times 1}$, as the vector for the number of recommended products that belong to each SITC 1-digit sector as a share of the total number of recommended export products.

The similarity score between the actual and the recommended export portfolios for country i at time t is then calculated as the similarity between the two vectors of actual and recommended structures:

$$Sim_{it} \equiv \frac{\left(S_{i,t-\Delta t}^{rec} - \bar{S}_{i,t-\Delta t}^{rec}\right) \cdot \left(S_{it}^{actual} - \bar{S}_{i,t}^{actual}\right)}{\|S_{i,t-\Delta t}^{rec} - \bar{S}_{i,t-\Delta t}^{rec}\| \|S_{it}^{actual} - \bar{S}_{i,t}^{actual}\|} \tag{2}$$

where Δt is a time lag we used when calculating the similarity scores to account for the fact that it takes time for an export structure to evolve.[12] This means that the similarity scores depend on our choices of time lags. In our baseline estimation, we set Δt = 5 years. Alternative assumptions for the time lags are also adopted in robustness checks presented in Section 6.

We calculate the annual Sim_{it} for all country-year pairs in the sample, and then incorporate the scores into the growth, volatility and risk-adjusted regressions that will be specified in the following section.

[12] We include this time lag because Che (2020) found that recommendations given by the product-based KNN algorithm are to some extent forward-looking, in that they match the export portfolios of several high-growth countries in their future years better than in the current year.

3.4 Growth and volatility estimations

Our main econometric exercise aims to investigate the correlation[13] of the alignment of recommended and actual export structure on growth and volatility of growth. The hypothesis is that countries with an export structure highly aligned with their latent comparative advantages– manifested as a high similarity score between their actual and recommended export portfolios, as defined in Section 3.3– should see higher and more stable growth over time.

To examine the impact of export structure on growth, we specify the following estimation model:

$$g_{it} = \beta_0 + \beta_1 y_{i,t-\Delta t} + \beta_2 Sim_{it} + \gamma \mathbf{X}_{it} + \epsilon_{it} \tag{3}$$

where g_{it} the average annual growth in GDP per capita for country i from $t-\Delta t$ to t. $y_{i,t-\Delta t}$ is the lagged real GDP per capita in log form. Sim_{it} is the similarity score calculated as in Section 3.3. $\mathbf{X}_{it}$ is a set of controls, including investment-to-GDP ratio, human capital growth, TFP growth, and world GDP growth in some specifications to account for common external shocks. We also include country and time fixed effects in $\mathbf{X}_{it}$ for some of the regression specifications (see Section 5). The similarity scores, as well as most control variables, are annual averages over the Δt time window. In our baseline estimation, we set Δt = 5 years. The regressions are run with non-overlapping Δt as the time unit. Our main parameter of interest is β_2.

Similarly, we can look at the impact of export structure on the volatility of growth with the following model:

$$vol_{it} = \beta_0 + \beta_1 vol_{i,t-\Delta t} + \beta_2 Sim_{it} + \gamma \mathbf{X}_{it} + \epsilon_{it} \tag{4}$$

where vol_{it} is the standard deviation of annual growth of real GDP per capita during the Δt time period. And $vol_{i,t-\Delta t}$ is the lagged dependent variable. Controls ($\mathbf{X}_{it}$) are broadly the same as in

[13] We are aware of the limit using the growth panel addressing causality. In the rest of the paper, "impact" is interpreted in correlation.

the growth regression, except we replace world growth with the growth volatility of world GDP, to control for the level of external volatility.

Alternatively, we can combine the information on the left-hand side of Equations 3 and 4, and estimate the impact of export structure on countries' "risk-adjusted growth". Here we define country i's risk-adjusted growth, g_{it}^{ra}, as the deviation of country i's growth from the world average growth rate, $g_{it} - \overline{g}_t$, divided by its standard deviation σ_{it}, over the Δt time period. We then estimate the following equation,

$$g_{it}^{ra} = \beta_0 + \beta_1 g_{i,t-\Delta t}^{ra} + \beta_2 Sim_{it} + \gamma \mathbf{X}_{it} + \epsilon_{it} \tag{5}$$

Controls ($\mathbf{X}_{it}$) are the same as in the growth regression, except we replace world growth with the average risk-adjusted growth across countries in $\mathbf{X}_{it}$.

For each equation, the estimation is done using simple OLS, country and time fixed-effect estimator, and a system GMM estimator following Arellano and Bond (1991). The system GMM estimator is employed to address the endogeneity issue introduced by having the lagged dependent variable on the right hand side, which likely affects the consistency of OLS and fixed-effect estimators. In the system GMM estimation, the lagged dependent variable and country-level controls are treated as endogenous and instrumented as such. Time fixed effect and world-level controls are treated as exogenous. Section 5 presents results from all three estimators for each regression.

4 Data

The country-product level raw export data and the actual export RCA scores come from the Atlas of Economic Complexity Dataverse (2020 version), which are in turn sourced from UN Comtrade Database. The macroeconomic variables come from the World Bank and Penn World Table[14]. Summary statistics for the main variables used in the regressions are shown in Table 1. The

[14]Version 10.0, see Feenstra et al. (2015) for metadata details.

data used for growth regressions are based on annual frequency covering 1980-2018. Specific estimations depend on different time-lag experiments as detailed in Section 5 and Section 6.

Table 1: Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Similarity score	0.817	0.179	-0.203	0.995	1203
GDP per capita	8.390	1.505	5.129	11.663	1414
Investment rate	0.219	0.109	-0.479	0.942	1398
TFP growth	0.002	0.028	-0.184	0.222	868
Human capital growth	0.01	0.007	-0.025	0.043	1107
GDP per capita growth	0.017	0.037	-0.247	0.367	1324
Growth volatility	0.015	0.014	0.001	0.142	1177
Risk-adjusted growth	0.705	4.474	-75.438	50.674	1175

The *similarity score* is calculated at the country-year level, following the steps described in Section 3.3. In the appendix, we show summary statistics for RCA scores and recommendation scores used to calculate the similarity score (see Table 18 and Table 19), as well as the box plots for the distributions of RCA scores and recommendation scores by SITC 1-digit sector (see Figure 10). Figure 8 presents the similarity score distribution around the world in 2018.

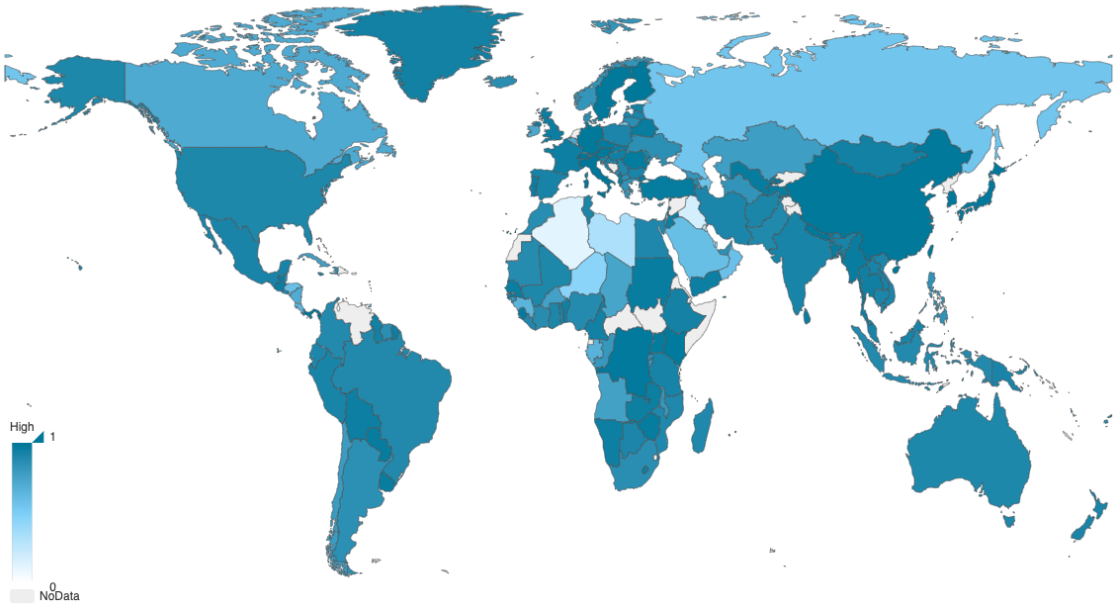


Figure 8: Similarity Score Distribution around the World

5 Estimation Results

5.1 Impact on growth

Table 2 presents our baseline results for the growth regression. Columns (1) and (2) are results from the OLS estimation with robust standard errors, without and with controls respectively. Column (3) presents results from the fixed-effect estimation with clustered standard errors. Columns (4) and (5) are results from the system GMM estimation, with year dummies and world GDP growth included in the controls, respectively. Standard errors are given in parentheses. As mentioned earlier, we choose a time window of 5 years in the baseline estimations, to account for the fact that export structure is a slow moving variable whose impact may take some time to show. This gives us over 700 observations in the GMM estimation.

As expected, the control variables– investment rate, TFP and human capital growth, world GDP growth– are positive and mostly significant. The lagged GDP per capita variable is negative and significant in the OLS and fixed effect estimations, consistent with the prediction of economic convergence theory that poorer countries should grow faster than richer countries. But the variable has a negative and insignificant coefficient in the system GMM specifications.

Consistent with our hypothesis, the similarity score variable is positive and significant in all regression specifications, i.e., higher alignment between actual export structure and recommended export structure is good for growth. The magnitude of the coefficient does not vary too much across different specifications. According to the system GMM estimation (Column 5), a 0.1 increase in the similarity score (Sim_{it}) is associated with a 0.22 percentage point increase in the annual growth rate of real GDP per capita. The coefficient is statistically significant at 1% level. This is equivalent to a move from the median to the 90th percentile of the similarity score distribution.

The choice of the time window potentially affects the magnitude and significance of the result.[15] We will explore alternative specifications for the time window in the robustness section

[15]Note that the default setting, with $\Delta t = 5$ years, is by our choice to start with considering the trade evolution and the panel data structure.

(Section 6).

5.2 Impact on growth volatility

Table 3 gives the results for the volatility regression, where the dependent variable is the standard deviation of annual growth during each Δt period. Similar to Table 2, Columns (1)-(3) are results from OLS and fixed effect estimations. And Columns (4)-(5) are system GMM results, with year dummies and volatility of world growth in the controls respectively.

The lagged volatility variable is positive and significant across all regression specifications, even in the fixed effect and system GMM specifications, where persistent and country specific volatility differences are supposed to be accounted for. This suggests significant stickiness in growth volatility. The control variables of investment rate, TFP growth, and human capital growth do not appear to have any material influence on the growth volatility, ceteris paribus. However, the dependent variable is shown to be highly correlated with the external environment, as indicated by the positive and significant coefficient for the variable of world growth volatility.

Turning to the similarity score, the results show a negative and significant coefficient for the variable under two OLS and two system GMM estimation specifications, i.e., higher alignment between actual export structure and recommended export structure helps growth to be more stable. The coefficient is negative but not significant in the fixed effect estimation.

In terms of the magnitude of impact, the system GMM estimator in Column 5 suggests that a 0.1 increase in the similarity score is associated with a 0.0015 decrease in the standard deviation of growth rate in a 5-year window. This is statistically significant at 1% level. In the robustness section, we will explore whether changing Δt alters the sign and significance of the baseline results.

5.3 Impact on risk-adjusted growth

Instead of looking at growth rate and growth volatility separately, we can also summarize a country's growth performance in one variable, which we call the risk-adjusted growth. This is calculated as the average annual growth divided by the standard deviation of growth during Δt period. Table 4 presents results for the risk-adjusted growth regression. The columns are structured similarly as in Tables 2 and 3.

The lagged dependent variable shows up with a positive coefficient in all regression specifications, though it is not significant except in the OLS specification without controls and the fixed effect specification. Among the control variables included, investment rate and TFP growth are shown to be associated with higher risk-adjusted growth, while the human capital variable is insignificant and its sign varying across different estimations. In addition, the dependent variable is strongly correlated with the risk adjusted growth at the world level.

Consistent with our hypothesis, the similarity score variable has a positive coefficient across all specifications. The coefficient is significant in the basic OLS and the system GMM estimations. The system GMM result (Column 5) suggests that a 0.1 increase in the similarity score is associated with 20.95 percentage points increase per standard deviation in the annual growth rate. To put it in an alternative way, an increase of 0.8 in the similarity score moves a country from the world medium in the risk-adjusted growth spectrum to the 75th percentile level. The result is statistically significant at 5% level.

6 Robustness

6.1 Changing time interval

In the baseline regressions, we set $\Delta t = 5$ years. In this section, we examine how our results may change with different assumptions for Δt. Tables 5 to 7 show results with $\Delta t = 3, 5,$ and 7 years, under the system GMM specification with all controls. Tables 8 to 13 present the results with these alternative time interval assumptions for all estimation specifications.

As Table 5 shows, the signs remain the same and the magnitude and significance level for our variable of interest (Sim_{it}) differ slightly from the baseline case ($\Delta t = 5$). When $\Delta t = 3$, a 0.1 increase in the similarity score is associated with 0.19 percentage point increase in growth, while this coefficient is 0.13 when $\Delta t = 7$. The results for the cases of $\Delta t = 3$ and 7 under OLS and fixed effect specifications can be found in Tables 8 and 9.

Table 6 summarizes results for the volatility regression for different time windows, under the system GMM specification. It shows that a 0.1 increase in the similarity score is associated with 0.0017 and 0.0013 decrease in the standard deviation of growth when $\Delta t = 3$ and 7 respectively. These are mostly consistent with the baseline result. For the estimates of OLS and fixed effect specifications, see Tables 10 and 11.

The adjusted-growth regression results with alternative time windows using the system GMM estimator are summarized in Table 7. And Tables 12 and 13 present the results for other regression specifications. As Tables 7 shows, when $\Delta t = 3$, a 0.1 increase in the similarity score is associated with 21.84 percentage point increase in the risk-adjusted annual growth rate, at 1% significance level. However, this value decreases to 13.79 when $\Delta t = 7$, and it is not statistically significant.[16]

A summary of estimates with time intervals varying between 3 to 7 can be found in Figure 9. In this plot, we show confidence intervals for the key coefficient in growth regressions, volatility regressions, and risk-adjusted growth regressions against consecutive time intervals.

6.2 Winsorization

We also run a set of regressions where we winsorize the dependent variables by removing the top and bottom 5% from the sample observations. Overall, signs and significance of the similarity score variable do not change markedly from the baseline results.

Table 14 presents results of the growth regression with the real GDP growth rate winsorized. The coefficient for the similarity score remains positive and significant across all specifications.

[16] As a further robustness check, we tried $\Delta t = 8$ and this value is 19.89 percentage point at 5% significance level. Specifically, the estimated coefficient for similarity score in the risk-adjusted growth regression under the system GMM specification is 1.989 with standard error 0.959.

In fact, the magnitude of the coefficient is slightly lower in four out of the five specifications compared to the baseline. According to the system GMM estimates (Column 5), a 0.1 increase in the similarity score (Sim) is associated with a 0.18 percentage point increase in real GDP per capita growth.

Table 15 show results of the winsorized volatility regression. Overall the estimates for the similarity score variable are somewhat weaker than in the baseline in terms of magnitudes, but still point to a negative impact of the similarity score on growth volatility. The coefficient for Sim_{it} is negative in all five specifications, but not significant in four out of five specifications.

Table 16 are results for the winsorized adjusted-growth regression. Similar to the baseline results, the coefficient of similarity score is positive and significant in the system GMM with the world growth control, though the magnitudes of the coefficient is smaller than in the baseline. According to the system GMM estimates (Column 5), a 0.1 increase in the similarity score is associated with 9.8 percentage points increase per standard deviation in the annual growth rate. This is much smaller than 20.95 percentage points increase in the baseline result. The statistical significance level changes from 1% to 10%. But overall, the winsorized results still confirm a positive correlation between the similarity score and countries' comprehensive growth performance.

7 Conclusion

One of the frequently voiced complaints from economists and policy makers regarding the use of machine learning algorithms in empirical studies is the seeming opaque nature of the algorithms. The human cognitive system can differentiate a picture of a dog from that of a cat easily. But there is very little theory, i.e., a linear and logical explanation, about why such discernment can be reliably made. Many machine learning algorithms share the same characteristics. These algorithms can be very effective in making realistic pattern-recognition judgments, but an articulated rationale of such judgements is often lacking. On the other hand, traditional parametric econometric studies are under-pined by economic theories with easy-to-understand trains of thought. But

the typical linear regression models are drastic simplifications of reality, which may reduce their usefulness in guiding practical decisions in more complex scenarios.

In this paper, we try to combine the merits of both worlds to shed some light on the importance of export structure evolution in the growth and income convergence process. We leveraged machine learning methodology to characterize the complex patterns in countries' latent comparative advantages and create export recommendations accordingly. We then use a standard linear regression model to evaluate the soundness of these recommendations by asking whether a country's growth performance is better if they had de facto "followed" these recommendations.

Specifically, we used a product-based KNN algorithm to provide annual export product recommendations at the SITC 4-digit level for over 190 economies, from 1980 to 2018. We then look at whether more alignment between a country's recommended export structure and its actual export structure has any impact on growth and growth volatility.

Our results confirm the merits of such algorithm-based export recommendations. They show that economies with a higher *similarity score* between recommended and actual export portfolios achieve better growth performance. In our baseline estimation, a 0.5 increase in the similarity score is associated with a 1.1 percentage point increase in the annual growth of real GDP per capita, and a 0.0075 decrease in the standard deviation of growth rate over 5-year time windows. These results are overall robust with respect to changing time intervals and removing outliers.

It is worth noting that though we believe the algorithm-produced export recommendations can be a useful tool for policymakers to evaluate industrial policy options and for private investors entering new markets, they are no substitutes for detailed and multidimensional analyses of the viability of any industry in a country. In addition, it goes without saying that knowing which industries a country may have comparative advantages in does not automatically translate into specific policy recommendations. Neither are we advocating for direct policy interventions in shaping a country's export structure. How a country can best support the growth of its tradable sector to leverage the country's comparative advantages is likely a case-by-case discussion and depends on many country-specific factors. We have considered this in the methodology design

and we acknowledge that specific institutional challenges vary significantly across the world.

Nonetheless, our paper sheds light on direction of potential reforms and ways to strengthen the economy. When considering higher level of accession to the global market, small open economies with weak fundamentals confront tradeoffs to what products to diversify into. Large emerging markets could have challenges associated with either short-term or long-term bottlenecks. Our paper can serve as a diagnostic tool for policymakers to consider broader reform agenda and specific export product diversification strategies. An example of this practical application can be found in IMF (2021).

We believe that knowing what a country's potential export portfolios may look like, comparing it with the reality, and investigating what may have caused the difference is a valuable exercise for policy makers to identify potential policy gaps and reform areas to focus on for achieving better growth.

References

Adeniyi, David Adedayo, Zhaoqiang Wei, and Y Yongquan. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method". In: *Applied Computing and Informatics* 12.1 (2016), pp. 90–108.

Aiginger, Karl and Dani Rodrik. "Rebirth of Industrial Policy and an Agenda for the Twenty-First Century". In: *Journal of Industry, Competition and Trade* (2020), pp. 1–19.

Arellano, Manuel and Stephen Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". In: *Review of Economic Studies* 58.2 (1991), pp. 277–297.

Bacha, Edmar L and Albert Fishlow. "The recent commodity price boom and Latin American growth: More than new bottles for an old wine?" In: *The Oxford Handbook of Latin American Economics*. 2011.

Balaguer, Jacint and Manuel Cantavella-Jorda. "Structural change in exports and economic growth: cointegration and causality analysis for Spain (1961–2000)". In: *Applied Economics* 36.5 (2004), pp. 473–477.

Balassa, Bela. "Trade Liberalisation and "Revealed" Comparative Advantage". In: *The Machester School* 33.2 (1965), pp. 99–123.

Cadot, Olivier, Céline Carrère, and Vanessa Strauss-Kahn. "Export diversification: what's behind the hump?" In: *Review of Economics and Statistics* 93.2 (2011), pp. 590–605.

Che, Natasha. "Intelligent Export Diversification: An Export Recommendation System with Machine Learning". In: *IMF Working Papers* 2020.175 (2020).

Clickstreams, Multi-faceted Web. "Workshop Notes". In: (2005).

Feenstra, Robert and Hiau Looi Kee. "Export variety and country productivity: Estimating the monopolistic competition model with endogenous productivity". In: *Journal of international Economics* 74.2 (2008), pp. 500–518.

Frankel, Jeffrey A. *The natural resource curse: a survey*. Tech. rep. National Bureau of Economic Research, 2010.

Giri, Rahul, Mr Saad N Quayyum, and Rujun Yin. *Understanding Export Diversification: Key Drivers and Policy Implications*. International Monetary Fund, 2019.

Hausmann, Ricardo, Jason Hwang, and Dani Rodrik. "What you export matters". In: *Journal of economic growth* 12.1 (2007), pp. 1–25.

Hausmann, Ricardo and Bailey Klinger. "The structure of the product space and the evolution of comparative advantage". In: *CID Working Paper Series* (2007).

Herzer, Dierk and Felicitas Nowak-Lehnmann D. "What does export diversification do for growth? An econometric analysis". In: *Applied economics* 38.15 (2006), pp. 1825–1838.

Hidalgo, César A and Ricardo Hausmann. "The building blocks of economic complexity". In: *Proceedings of the national academy of sciences* 106.26 (2009), pp. 10570–10575.

Imbs, Jean and Romain Wacziarg. "Stages of diversification". In: *American Economic Review* 93.1 (2003), pp. 63–86.

IMF. "Sustaining Long-Run Growth and Macroeconomic Stability in Low-Income Countries-The Role of Structural Transformation and Diversification". In: *IMF Policy Paper* (2014).

— "Uruguay: Selected Issues". In: *IMF Country Report* (2021).

Klinger, Bailey and Daniel Lederman. *Discovery and development: An empirical exploration of "new" products*. The World Bank, 2004.

— "Export discoveries, diversification and barriers to entry". In: *Economic Systems* 35.1 (2011), pp. 64–83.

Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems". In: *Computer* 42.8 (2009), pp. 30–37.

Lathia, Neal, Stephen Hailes, and Licia Capra. "kNN CF: a temporal social network". In: *Proceedings of the 2008 ACM conference on Recommender systems*. 2008, pp. 227–234.

Lin, Justin Yifu and Feiyue Li. *Development strategy, viability, and economic distortions in developing countries*. The World Bank, 2009.

Al-Marhubi, Fahim. "Export diversification and growth: an empirical investigation". In: *Applied economics letters* 7.9 (2000), pp. 559–562.

Paterek, Arkadiusz. “Improving regularized singular value decomposition for collaborative filtering”. In: *Proceedings of KDD cup and workshop*. Vol. 2007. 2007, pp. 5–8.

Prebisch, Raul. “The economic development of Latin America and its principal problems”. In: *Economic Bulletin for Latin America* (1962).

Sarwar, Badrul et al. *Application of dimensionality reduction in recommender system-a case study*. Tech. rep. Minnesota Univ Minneapolis Dept of Computer Science, 2000.

— “Incremental singular value decomposition algorithms for highly scalable recommender systems”. In: Citeseer.

Singer, Hans W. “The distribution of gains between investing and borrowing countries”. In: *The Strategy of International Development*. Springer, 1975, pp. 43–57.

Singer, HW. “The Distribution of Gains between Investing and Borrowing Countries”. In: *The American Economic Review* 40.2 (1950), pp. 473–485.

Table 2: Growth Regression (baseline)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged GDP Per Capita	-0.002**	-0.008***	-0.016***	-0.001	-0.001
	(0.001)	(0.003)	(0.004)	(0.001)	(0.001)
Similarity Score	0.023***	0.036***	0.035**	0.023***	0.022***
	(0.008)	(0.011)	(0.014)	(0.006)	(0.006)
Inv. Rate		0.016***	0.014***	0.021***	0.021***
		(0.002)	(0.002)	(0.004)	(0.004)
TFP Growth		0.716***	0.689***	0.745***	0.742***
		(0.063)	(0.057)	(0.064)	(0.061)
Human Capital Growth		0.426***	0.420***	0.600***	0.571***
		(0.127)	(0.133)	(0.149)	(0.153)
World Growth		0.381*			0.263
		(0.195)			(0.193)
No. of Obs.	1168	732	732	732	732
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.01	0.02
Hansen-J (p-value)				0.79	0.80

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Volatility Regression (baseline)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Volatility	0.443***	0.272***	0.268***	0.468***	0.468***
	(0.071)	(0.075)	(0.059)	(0.056)	(0.055)
Similarity Score	-0.010***	-0.023**	-0.022	-0.014**	-0.015***
	(0.004)	(0.011)	(0.014)	(0.006)	(0.006)
Inv. Rate		-0.001	-0.001	-0.003**	-0.002
		(0.001)	(0.001)	(0.001)	(0.001)
TFP Growth		-0.011	-0.013	-0.012	-0.011
		(0.035)	(0.026)	(0.023)	(0.027)
Human Capital Growth		0.136*	0.138	0.112	0.082
		(0.080)	(0.090)	(0.083)	(0.078)
World Volatility		0.675***			0.464**
		(0.190)			(0.209)
No. of Obs.	966	614	613	614	614
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.51	0.53
Hansen-J (p-value)				0.58	0.38

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Risk-adjusted Growth Regression (baseline)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Adj. Growth	0.311***	0.073	0.086*	0.066	0.050
	(0.111)	(0.045)	(0.048)	(0.052)	(0.046)
Similarity Score	3.406***	3.270	2.604	1.985**	2.095**
	(0.627)	(2.589)	(2.568)	(0.863)	(0.853)
Inv. Rate		1.989***	2.247***	1.775***	1.721***
		(0.487)	(0.527)	(0.482)	(0.541)
TFP Growth		33.794***	38.395***	36.424***	29.998***
		(8.458)	(11.167)	(10.365)	(8.329)
Human Capital Growth		1.194	-11.629	-25.097	-14.007
		(29.750)	(28.030)	(34.874)	(34.170)
World Adj. Growth		0.508**			0.578***
		(0.218)			(0.215)
No. of Obs.	964	613	612	613	613
AR1 (p-value)				0.01	0.02
AR2 (p-value)				0.49	0.17
Hansen-J (p-value)				0.49	0.43

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Robustness: Growth

	(1) +3	(2) +5	(3) +7
Lagged GDP Per Capita	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.002)
Similarity Score	0.019***	0.022***	0.013**
	(0.007)	(0.006)	(0.007)
Inv. Rate	0.021***	0.021***	0.018***
	(0.003)	(0.004)	(0.004)
TFP Growth	0.797***	0.742***	0.758***
	(0.067)	(0.061)	(0.076)
Human Capital Growth	0.650***	0.571***	0.530***
	(0.158)	(0.153)	(0.183)
World Growth	0.391***	0.263	0.613***
	(0.085)	(0.193)	(0.210)
Constant	0.034**	0.030**	0.030
	(0.015)	(0.015)	(0.019)
No. of Obs.	1258	732	523
AR1 (p-value)	0.00	0.00	0.01
AR2 (p-value)	0.89	0.02	0.10
Hansen-J (p-value)	1.00	0.80	0.01

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Robustness: Volatility

	(1) +3	(2) +5	(3) +7
Lagged Volatility	0.417***	0.468***	0.243***
	(0.058)	(0.055)	(0.073)
Similarity Score	-0.017***	-0.015***	-0.013***
	(0.005)	(0.006)	(0.004)
Inv. Rate	-0.002	-0.002	-0.002
	(0.001)	(0.001)	(0.001)
TFP Growth	-0.116***	-0.011	0.088**
	(0.032)	(0.027)	(0.035)
Human Capital Growth	0.075	0.082	-0.058
	(0.102)	(0.078)	(0.102)
World Volatility	0.526***	0.464**	0.532**
	(0.132)	(0.209)	(0.234)
Constant	0.010*	0.008*	0.009*
	(0.006)	(0.005)	(0.005)
No. of Obs.	1231	614	409
AR1 (p-value)	0.00	0.00	0.00
AR2 (p-value)	0.05	0.53	0.31
Hansen-J (p-value)	1.00	0.38	0.23

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Robustness: Risk-adjusted Growth

	(1) +3	(2) +5	(3) +7
Lagged Adj. Growth	0.265***	0.050	0.253***
	(0.055)	(0.046)	(0.059)
Similarity Score	2.184***	2.095**	1.379
	(0.679)	(0.853)	(0.893)
Inv. Rate	1.483***	1.721***	0.442
	(0.320)	(0.541)	(0.400)
TFP Growth	36.657***	29.998***	58.120***
	(7.757)	(8.329)	(11.848)
Human Capital Growth	49.083**	-14.007	73.046*
	(21.635)	(34.170)	(40.394)
World Adj. Growth	0.528***	0.578***	0.715***
	(0.156)	(0.215)	(0.164)
Constant	0.358	1.486	-1.217
	(0.703)	(1.020)	(0.900)
No. of Obs.	1230	613	408
AR1 (p-value)	0.00	0.02	0.00
AR2 (p-value)	0.14	0.17	0.01
Hansen-J (p-value)	1.00	0.43	0.10

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Growth Regression (fwd+3)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged GDP Per Capita	-0.001*	-0.008***	-0.015***	-0.002	-0.001
	(0.001)	(0.002)	(0.005)	(0.001)	(0.001)
Similarity Score	0.025***	0.022**	0.023**	0.020***	0.019***
	(0.007)	(0.010)	(0.010)	(0.007)	(0.007)
Inv. Rate		0.019***	0.018***	0.022***	0.021***
		(0.002)	(0.003)	(0.004)	(0.003)
TFP Growth		0.764***	0.740***	0.779***	0.797***
		(0.050)	(0.069)	(0.070)	(0.067)
Human Capital Growth		0.492***	0.486***	0.566***	0.650***
		(0.121)	(0.167)	(0.196)	(0.158)
World Growth		0.422***			0.391***
		(0.101)			(0.085)
No. of Obs.	2030	1258	1258	1258	1258
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.95	0.89
Hansen-J (p-value)				1.00	1.00

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Growth Regression (fwd+7)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged GDP Per Capita	-0.002**	-0.011***	-0.024***	-0.001	-0.001
	(0.001)	(0.003)	(0.006)	(0.002)	(0.002)
Similarity Score	0.015**	0.019	0.021**	0.014**	0.013**
	(0.007)	(0.015)	(0.010)	(0.007)	(0.007)
Inv. Rate		0.014***	0.012***	0.018***	0.018***
		(0.002)	(0.002)	(0.005)	(0.004)
TFP Growth		0.686***	0.622***	0.770***	0.758***
		(0.076)	(0.093)	(0.093)	(0.076)
Human Capital Growth		0.481***	0.496***	0.580**	0.530***
		(0.176)	(0.187)	(0.239)	(0.183)
World Growth		1.054***			0.613***
		(0.301)			(0.210)
No. of Obs.	838	523	522	523	523
AR1 (p-value)				0.01	0.01
AR2 (p-value)				0.13	0.10
Hansen-J (p-value)				0.01	0.01

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Volatility Regression (fwd+3)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Volatility	0.444***	0.304***	0.329***	0.450***	0.417***
	(0.044)	(0.061)	(0.055)	(0.054)	(0.058)
Similarity Score	-0.014***	-0.011	-0.011	-0.015**	-0.017***
	(0.003)	(0.009)	(0.011)	(0.006)	(0.005)
Inv. Rate		-0.001	-0.002	-0.003**	-0.002
		(0.001)	(0.001)	(0.002)	(0.001)
TFP Growth		-0.113***	-0.117***	-0.117***	-0.116***
		(0.031)	(0.029)	(0.033)	(0.032)
Human Capital Growth		0.025	0.059	0.084	0.075
		(0.089)	(0.108)	(0.121)	(0.102)
World Volatility		0.661***			0.526***
		(0.118)			(0.132)
No. of Obs.	1948	1231	1231	1231	1231
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.04	0.05
Hansen-J (p-value)				1.00	1.00

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Volatility Regression (fwd+7)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Volatility	0.453***	-0.013	-0.017	0.235***	0.243***
	(0.098)	(0.096)	(0.058)	(0.069)	(0.073)
Similarity Score	-0.009**	-0.016	-0.017	-0.013***	-0.013***
	(0.004)	(0.010)	(0.011)	(0.005)	(0.004)
Inv. Rate		-0.002	-0.002	-0.003*	-0.002
		(0.001)	(0.001)	(0.001)	(0.001)
TFP Growth		0.065	0.068*	0.102***	0.088**
		(0.039)	(0.038)	(0.033)	(0.035)
Human Capital Growth		0.167	0.129	-0.121	-0.058
		(0.123)	(0.151)	(0.098)	(0.102)
World Volatility		0.908***			0.532**
		(0.247)			(0.234)
No. of Obs.	645	409	402	409	409
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.25	0.31
Hansen-J (p-value)				0.30	0.23

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Risk-adjusted Growth Regression (fwd+3)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Adj. Growth	0.407***	0.156***	0.164***	0.280***	0.265***
	(0.050)	(0.043)	(0.038)	(0.056)	(0.055)
Similarity Score	2.641***	2.259*	2.870**	2.403***	2.184***
	(0.418)	(1.167)	(1.311)	(0.824)	(0.679)
Inv. Rate		2.084***	2.257***	1.520***	1.483***
		(0.322)	(0.321)	(0.318)	(0.320)
TFP Growth		35.252***	37.052***	40.304***	36.657***
		(4.555)	(7.304)	(7.740)	(7.757)
Human Capital Growth		57.144***	38.695**	31.331	49.083**
		(19.311)	(18.272)	(24.025)	(21.635)
World Adj. Growth		0.410***			0.528***
		(0.137)			(0.156)
No. of Obs.	1946	1230	1230	1230	1230
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.16	0.14
Hansen-J (p-value)				1.00	1.00

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Risk-adjusted Growth Regression (fwd+7)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Adj. Growth	0.234**	0.137**	0.153***	0.289***	0.253***
	(0.107)	(0.054)	(0.045)	(0.060)	(0.059)
Similarity Score	3.283***	1.676	0.965	1.189	1.379
	(0.680)	(2.548)	(2.452)	(0.993)	(0.893)
Inv. Rate		1.251***	1.321***	0.457	0.442
		(0.446)	(0.448)	(0.393)	(0.400)
TFP Growth		44.795***	49.179***	65.872***	58.120***
		(11.965)	(15.320)	(13.691)	(11.848)
Human Capital Growth		43.964	25.323	58.270	73.046*
		(53.825)	(55.088)	(42.408)	(40.394)
World Adj. Growth		0.732***			0.715***
		(0.191)			(0.164)
No. of Obs.	643	408	401	408	408
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.03	0.01
Hansen-J (p-value)				0.13	0.10

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Growth Regression (winsorized)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged GDP Per Capita	-0.000	-0.008***	-0.018***	-0.002**	-0.002**
	(0.000)	(0.003)	(0.006)	(0.001)	(0.001)
Similarity Score	0.023***	0.019*	0.018*	0.018***	0.018***
	(0.006)	(0.010)	(0.009)	(0.007)	(0.007)
Inv. Rate		0.017***	0.015***	0.018***	0.018***
		(0.002)	(0.002)	(0.003)	(0.003)
TFP Growth		0.676***	0.651***	0.695***	0.693***
		(0.062)	(0.074)	(0.075)	(0.071)
Human Capital Growth		0.325***	0.306**	0.391***	0.367**
		(0.117)	(0.120)	(0.138)	(0.147)
World Growth		0.339**			0.313*
		(0.168)			(0.165)
Constant	0.001	0.076***	0.177***	0.048***	0.041***
	(0.007)	(0.025)	(0.049)	(0.013)	(0.013)
No. of Obs.	931	631	627	631	631
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.04	0.11
Hansen-J (p-value)				0.85	0.81

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Volatility Regression (winsorized)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Volatility	0.329***	0.141**	0.129***	0.311***	0.300***
	(0.044)	(0.069)	(0.043)	(0.038)	(0.038)
Similarity Score	-0.004*	-0.005	-0.005	-0.008	-0.007
	(0.002)	(0.006)	(0.006)	(0.005)	(0.005)
Inv. Rate		-0.002	-0.002	-0.002**	-0.002
		(0.001)	(0.001)	(0.001)	(0.001)
TFP Growth		-0.011	-0.014	0.001	0.002
		(0.029)	(0.022)	(0.023)	(0.024)
Human Capital Growth		0.029	0.026	-0.023	-0.034
		(0.056)	(0.054)	(0.057)	(0.058)
World Volatility		0.560***			0.449***
		(0.147)			(0.160)
Constant	0.011***	0.014*	0.011*	0.010*	0.004
	(0.002)	(0.007)	(0.006)	(0.005)	(0.005)
No. of Obs.	816	530	524	530	530
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.32	0.29
Hansen-J (p-value)				0.24	0.26

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Risk-adjusted Growth Regression (winsorized)

	(1) OLS	(2) OLS-X	(3) FE(t),X	(4) GMM-(t)	(5) GMM-X(t)
Lagged Adj. Growth	0.435***	0.225***	0.274***	0.388***	0.313***
	(0.038)	(0.051)	(0.046)	(0.058)	(0.058)
Similarity Score	1.878***	0.478	-0.181	0.699	0.980*
	(0.403)	(1.661)	(1.654)	(0.503)	(0.507)
Inv. Rate		0.850***	1.083***	0.467*	0.245
		(0.268)	(0.288)	(0.273)	(0.295)
TFP Growth		28.479***	31.456***	37.499***	32.498***
		(6.663)	(8.919)	(8.468)	(6.919)
Human Capital Growth		22.874	13.397	20.382	31.930
		(18.292)	(16.395)	(25.115)	(25.516)
World Adj. Growth		0.374***			0.438***
		(0.110)			(0.113)
Constant	-1.136***	0.502	2.163	-0.287	-0.680
	(0.331)	(0.666)	(1.588)	(0.731)	(0.710)
No. of Obs.	804	522	517	522	522
AR1 (p-value)				0.00	0.00
AR2 (p-value)				0.81	0.22
Hansen-J (p-value)				0.23	0.26

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 17: Classification according to the SITC 1 -Section

SITC Code	**Sector Name**
0	Food and live animals
1	Beverages and tobacco
2	Crude materials, inedible, except fuels
3	Mineral fuels, lubricants and related materials
4	Animal and vegetable oils, fats and waxes
5	Chemicals and related products, n.e.s.
6	Manufactured goods classified chiefly by material
7	Machinery and transport equipment
8	Miscellaneous manufactured articles
9	Commodities and transactions not classified elsewhere in the SITC

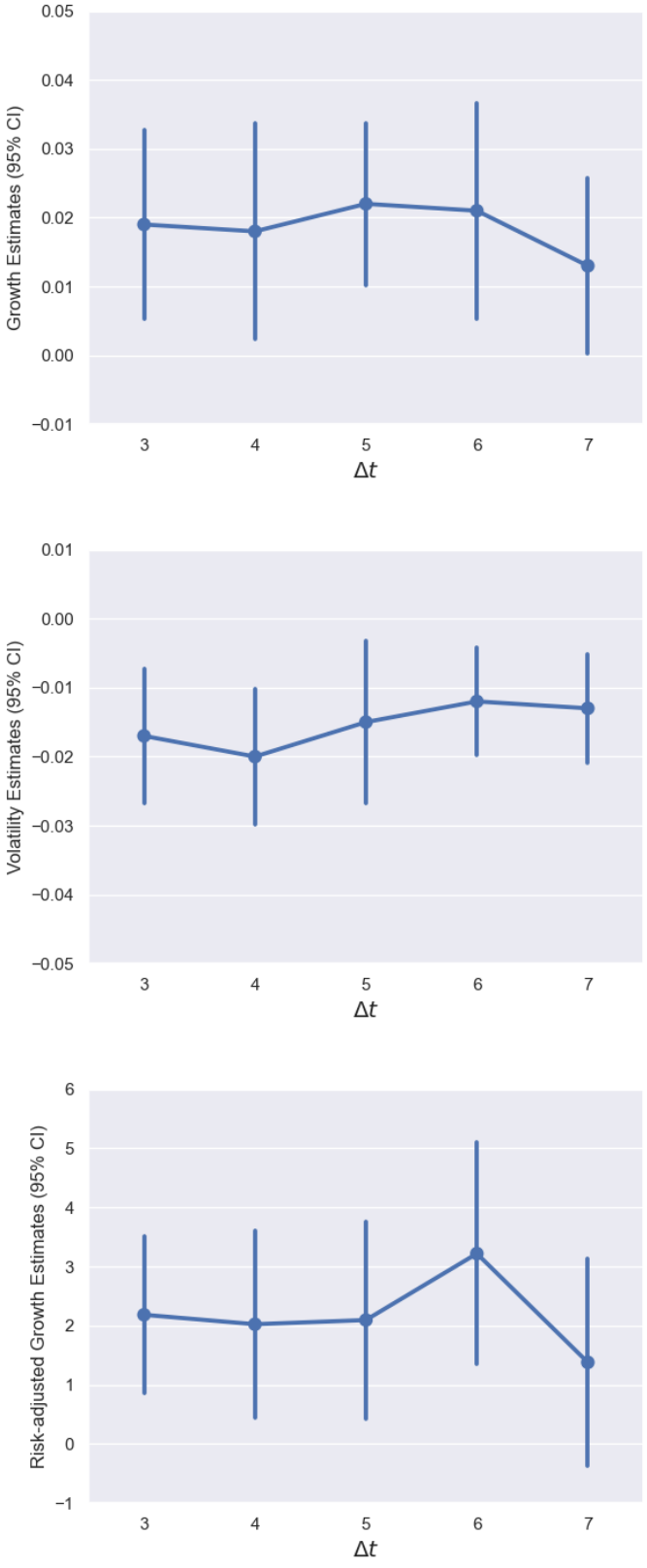


Figure 9: Robustness: Confidence Intervals in Main Regressions

Table 18: Summary of Actual RCA by year

year	N	Mean	Q1	Median	Q3
1985	61160	0.136	0.024	0.158	0.844
1990	69130	0.133	0.024	0.156	0.788
1995	84213	0.144	0.027	0.168	0.798
2000	96925	0.138	0.027	0.163	0.780
2005	100931	0.128	0.024	0.155	0.744
2010	104069	0.120	0.023	0.152	0.735
2015	101789	0.117	0.022	0.143	0.708

Table 19: Summary of Recommendation Scores by year

year	N	Mean	Q1	Median	Q3
1985	133722	0.269	0.140	0.328	0.648
1990	134504	0.308	0.161	0.370	0.686
1995	152281	0.350	0.202	0.403	0.712
2000	159444	0.365	0.217	0.438	0.754
2005	160218	0.348	0.198	0.419	0.749
2010	160784	0.342	0.195	0.416	0.747
2015	160576	0.337	0.194	0.393	0.702

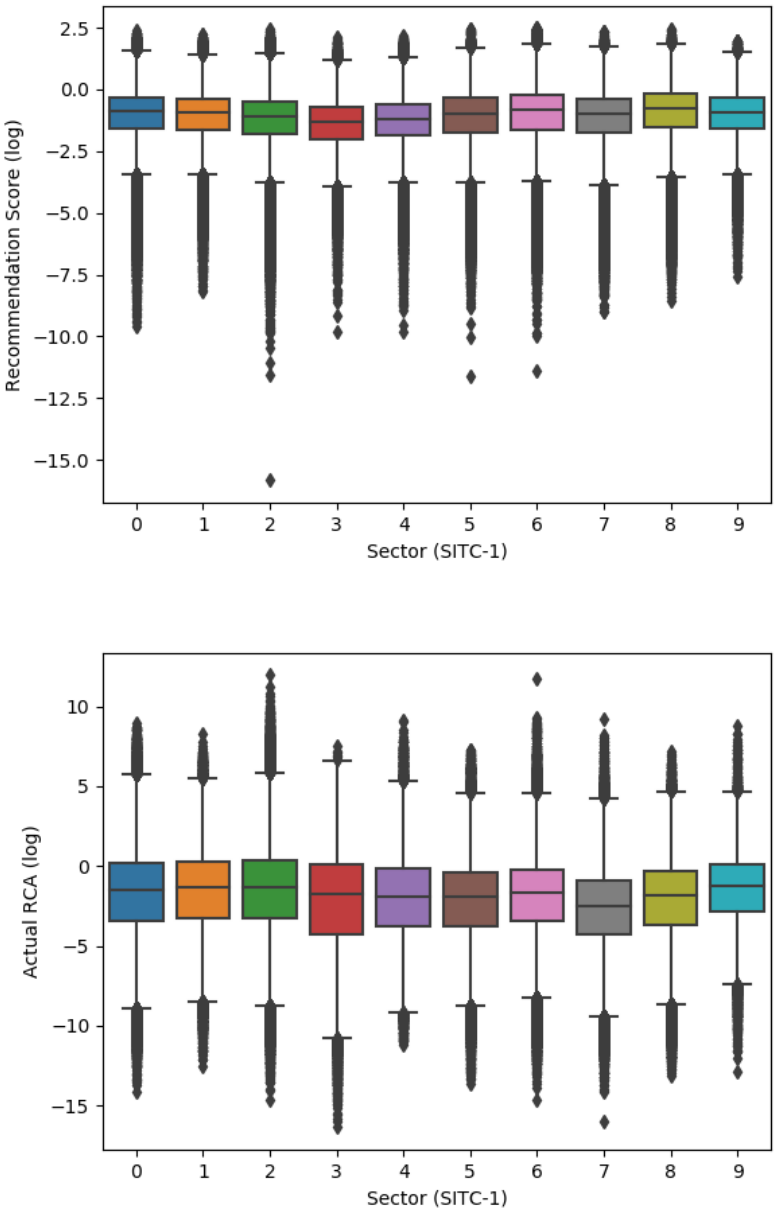


Figure 10: Distribution of Actual RCA and Recommendation Scores by Sector

PUBLICATIONS

High Performance Export Portfolio: Design Growth-Enhancing Export Structure with Machine Learning

Working Paper No. WP/2022/075