Maastricht University

UNITED NATIONS UNIVERSITY
UNU-MERIT

Working Paper Series

#2022-015

The canonical correlation complexity method

Önder Nomaler & Bart Verspagen

Published 20 April 2022

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

The Canonical Correlation Complexity Method

Önder Nomaler & Bart Verspagen

UNU-MERIT

Abstract:

A relatively recent, yet rapidly proliferating strand of literature in the so-called econophysics domain, known as 'economic complexity' , introduces a toolkit to analyse the relationship between specialization, diversification, and economic development. Different methods that aim at reducing the high dimensionality in data on the empirical patterns of co-location (be it nations or regions) of specializations have been proposed. In terms of the concepts of machine learning, the existing algorithms follow the framework of 'unsupervised learning'. The competing alternatives (e.g., Hidalgo and Hausmann, 2009 vs. Tacchella et al, 2012) have been based on very different assessments of which products depend on more complex capabilities, and accordingly yield highly different estimations of complexity at the product level. The approach that we developed avoids this algorithmic 'confusion' by drawing on a toolkit of more transparent and long-established methods that follow the 'supervised learning' principle where the data on trade/specialization and development are processed together from the very beginning in order to identify the patterns of mutual association. The first pillar of the toolkit, Principal Component Analysis (PCA), serves dimensionality reduction in co-location information. The second pillar, Canonical Correlation Analysis (CCA), identifies the mutual-association between the various patterns of (co-)specialization and more-than-one dimension of economic development. This way, we are able to identify the products or technologies that can be associated with the level or the growth rate of per capita GDP and CO_2 emissions.

Keywords: Economic complexity; economic development; supervised learning; canonical correlation analysis; principal component analysis.

JEL codes: F14; F63; O11

20 April 2022

This paper is an outcome of a project funded by the Economic Complexity unit of the Joint Research Centre (JRC, Seville) of the European Commission through 2020-2021. We thank Emanuele Pugliese and participants at two workshops at the JRC, Seville for comments to earlier drafts of the paper. The views expressed and any remaining errors are the sole responsibility of the authors.

1. Introduction

Economic growth is a complex process in which there is not only an expansion of living standards, but also a deep structural change of the economy that is developing (see, e.g., Abramovitz, 1989, Fagerberg and Verspagen, 2010; Verspagen, 2005). This means that a developed economy not only has a much higher living standard than an underdeveloped economy, but that it also produces and consumes different products and services. This has been analyzed in a large literature on economic growth, development and structural change (see, e.g., Cantore and Alcorta, 2021; Nomaler and Verspagen, 2021, for overviews of this literature).

More recently, it has led to a developing literature in the so-called econophysics domain, which aims to develop new indicators related to the development process. The central idea here is that structural differences between countries will be visible in high-dimensional data on exports, and that these structural differences can be used as indicators for development. Freire (2021) surveys this literature. In this paper, we will propose a new method in this field, using the same high-dimensional export data, but a new algorithm that differs substantially from the dominant approaches.

Our algorithm distinguishes itself by the characteristic that is a supervised learning algorithm, i.e., it analyzes the export data with the explicit aim to maximize correlations to a set of indicators related to economic development (we call this the competitiveness indicators). As will be explained in some detail below, the existing approaches can be seen as unsupervised learning approaches, i.e., they derive indicators from the export data in isolation, and *ex post* try to relate these indicators to development. We feel that because structural change, growth and development are closely intertwined and co-evolutionary processes (Verspagen, 2005), the supervised learning approach that we propose is adequate as an attempt to summarize the main characteristics of development that can be found in the actual data.

As our method is yet undocumented, one main goal of this paper is to describe the method, which we call the Canonical Correlation Analysis Complexity method (or CCA complexity) in technical terms so that it can be reproduced. We also present some empirical results from the CCA complexity method, and compare these results to two algorithms that we proposed in the earlier literature.

The paper is structured as follows. In Section 2, we present some brief empirical results from the two existing methods. Our aim here is to show how these methods generate indicators both at the country level and at the export product level, how these indicators can be interpreted, and how they relate to each other. Interestingly (although not unexpected), we find that the indicators from both methods differ from each other. Section 2 concludes by some brief reflections about why the supervised learning method we propose may provide additional insights to the existing methods.

In Section 3, we present the details of our proposed method. This includes both an intuitive description that will be comprehensible for the statistics-savvy reader, and mathematical details that can be used to reproduce the method. A Matlab script that applies the method as described in Section 3 will be made available later.

The remaining sections are aimed at illustrating the method with actual data. We choose to use a pooled data sample for a wide sample of countries for the period 1996 – 2018 (this includes both

a sample to estimate the model, and data used to test predictions from it). The empirical illustration of our method proceeds as follows. In Section 4, we show how we select the value of an important parameter (*f*) in the method. This parameter regulates how much the high-dimensional trade data are reduced before they enter the main analysis. Section 4 concludes by specifying the particular value for *f* that we will use throughout the remaining sections.

Section 5 then presents the basic estimations results of the CCC method. Among other things, it documents the estimated coefficients, the goodness of fit measures (canonical correlations), the stability of the derived indicators over time, and a comparison of the indicators (and their stability) to the other methods found in the literature. Section 6 focuses on how well the method predicts the values of the variables in the competitiveness data set. Both in-sample and out-of-sample predictions are presented. The predictions of the CCC method are also compared to similar predictions from the other algorithms. In Section 7 we present some further reflections on how the set of three algorithms (our own CCC method and the two pre-existing methods) differ.

Finally, Section 8 summarizes the argument and presents the main conclusions.

2. Motivation

In this section, we will briefly review two existing methods of complexity analysis aimed at analyzing detailed exports data, and discuss which role canonical correlation analysis can play in adding to this existing toolbox. We will start with the method introduced by Hidalgo and Hausman (2009), which opened the field. Although this method has evolved since it was first published (Albeaik et al., 2017, 2017a), we will focus on the original version, because the additions to the method do not significantly change our interpretation of it. Second, we will review the method by Tacchella et al. (2012), which we will refer to as the Tea method. We will present empirical illustrations of both methods, and investigate how they relate to each other in terms of their assessment of the competitiveness of countries and the specific role of products in this.

After this, we will discuss how Canonical Correlation Analysis (CCA), in combination with PCA can add to the complexity studies. We will focus on two issues: how the results of the complexity methods can be interpreted, and the stability of the indicators that are derived.

2.1. A brief overview of alternative methods

Hidalgo and Hausman introduced what they called the method of reflections to operationalize complexity in detailed data on exports. They start from an *n*x*c* matrix **M** of so-called revealed comparative advantage indicators, where the (binary) elements are defined as

m_{ip} = 1 if $(X_{ip}/X_{\mathrm{i}})/(X_p/X)$ >1 and m_{ip} = 0 otherwise (2.1)

Here, X_{ip} is the export value of country *i* in product *p*, and the absence of a subscript points to aggregation over the relevant dimensions. The number of countries is denoted by *c*, and the number of products is denoted by *n*.

The matrix **M** is used to calculate the diversification (denoted by *D*) of country *i* and the ubiquity (denoted by *U*) of product *j*, as follows:

$$D_i = \sum_p M_{ip} \, , U_j = \sum_c M_{cp} \quad (2.2)$$

The basic idea underlying the method is that countries need specific production capabilities to start producing new products. Thus, increasing diversification is a process of accumulating production capabilities. However, capabilities differ between products, with some products requiring advanced capabilities, and others less advanced. Ubiquity is taken as an indicator for the level of production capabilities associated with the product: products with low (high) ubiquity require (less) advanced production capabilities.

Hidalgo and Hausman then propose an iterative procedure, in which successive 'reflections' generate new indicators at the country and product level. *D* and *U* are the zero-level iterations:

$$k_{i0}^{c} = D_i, k_{j0}^{p} = U_j \tag{2.3}$$

The iterations *N* (with *N*>0) proceed as follows:

$$k_{iN}^{c} = \frac{1}{D_i}\sum_q M_{iq} k_{qN-1}^{p}, k_{jN}^{p} = \frac{1}{U_j}\sum_q M_{qj} k_{qN-1}^{c} \tag{2.4}$$

Thus, the country-indicator derived in the *N*th iteration is equal to the sum of the previous-iteration product-indicator of the products in which the country has a comparative advantage, and the *N*th iteration product-indicator is equal to the sum of previous-iteration country-indicator of all countries that have a comparative advantage in the product. One interpretation of this iterative procedure is that with every iteration, we get closer to the interplay between product-level capabilities and country-level development level (the achieved set of production capabilities) that simultaneously determine the observed levels of *U* and *D*.

The nature of the resulting indicators differs between odd and even iterations: "For countries, even variables … are generalized measures of diversification, whereas odd variables … are generalized measures of the ubiquity of their exports. For products, even variables are related to their ubiquity and the ubiquity of other related products, whereas odd variables are related to the diversification of countries exporting those products." (Hidalgo and Hausman, 2009, p. 10571). We will focus on the even iterations here. Hidalgo and Hausman (p. 10574) argue that the even iterations for the country indicator provides "… information [that] … is related to factors that affect the ability to generate per capita income." The even iterations of the product indicator then show how each of the products contribute to this ability to generate income.

The Tea method follows some of the basic principles of HH, but differs on a few others. Like HH, it is an iterative method. The two main differences relative to HH are, first, that "the complexity of a product is *inversely* proportional to the number of countries which export it" (Tacchella et al., 2012, p. 2; our emphasis). While HH also recognize that complexity of a product is inversely related to ubiquity, their calculations (as in equations 2.3 and 2.4 above), their calculations do not reflect this explicitly (i.e., they leave this to the interpretation of their reflections). Second, the Tea method normalizes the country and product indicator at each step. This is not part of the HH method.

The equations for the Tea method are then as follows.

$$F_{i0} = 1, Q_{j0} = 1 \tag{2.5}$$

$$\tilde{F}_{iN} = \sum_q M_{iq} Q_{qN-1}, \tilde{Q}_{jN} = \frac{1}{\sum_q M_{qj} \frac{1}{F_{qN-1}}} \tag{2.6}$$

$$F_{iN} = \frac{\tilde{F}_{iN}}{\langle \tilde{F}_{pN} \rangle_p}, Q_{iN} = \frac{\tilde{Q}_{iN}}{\langle \tilde{Q}_{qN} \rangle_q} \quad (2.7)$$

where F_{iN} is the country indicator (comparable to k_{iN}^{c} from HH) and Q_{jN} is the product indicator (comparable to k_{jN}^{p}). Equation (2.7) is the normalization part of the method, and in this equation $\langle \quad \rangle_z$ indicates an average over all z entities (countries or products).

2.2. Empirical illustration

We will briefly illustrate these two methods for data of the year 1998. The reason for going so far back in time is that in this way, we will be able to relate the indicators to 20-year growth rates, which we consider the very long run. The ultimate source of our trade data at the product level is the United Nations' COMTRADE database, but we use the database as processed by Tacchella (2020). From the RCA indicators presented in this paper, we select the raw, unworked variety (tRCA) for this stage of the analysis (when we present the data for our own method at a later stage, we will switch to a different indicator), because this provides the largest similarity to the initial results presented by the authors of the methods that we outlined above.[1] For GDP per capita, we use data from the World Bank's World Development Indicators database, and we measure this variable in constant 2011 PPP international dollars.

In case of the method of reflections, we calculated reflections 0-20. We will present the even reflections 0-8 here, beyond reflection 8, every additional reflection correlated almost perfectly with the previous one (R>0.99). We also follow Hidalgo and Hausman by analyzing the statistical relationship of the (even) reflections as well as the Tea fitness indicator to GDP per capita and its growth rate. We will also summarize the distribution of the product indicator at successive reflections, and compare to the Tea product quality indicator. The database used for this purpose contains 152 countries and 5,035 products in HS1996. However, not all countries have comparative advantage in at least one product, and not all products have at least one country with comparative advantage. Such products or countries are dropped from the analysis.

[1] Note that like in the original versions of HH and Tea, we use a binarized variety of (t)RCA as in equation 2.2 above. The raw version is simply $(X_{ip}/X_i)/(X_p/X)$.

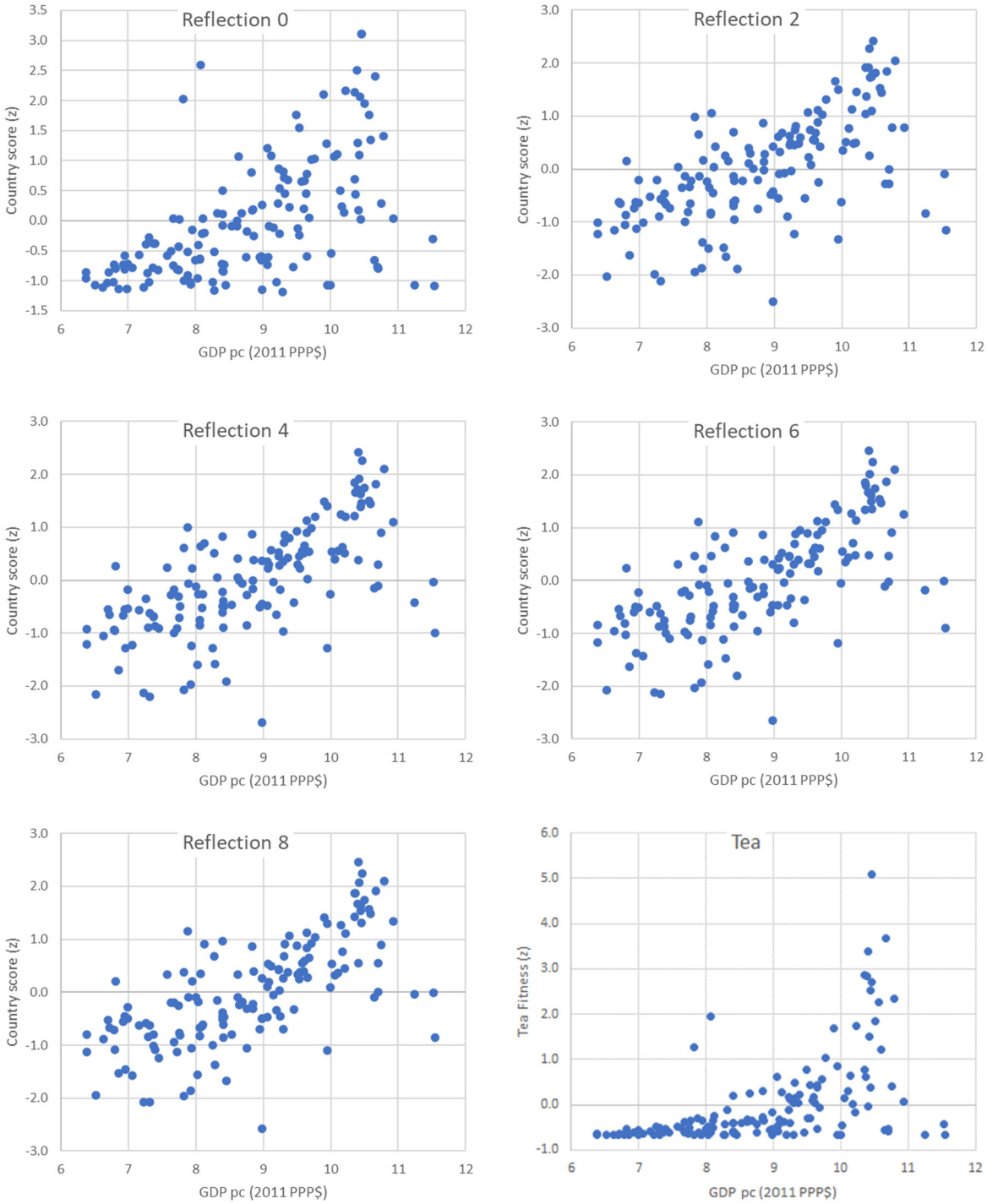


Figure 2.1. Even HH reflections 0-8 and Tea Fitness graphed against GDP per capita

Figure 2.1. shows the correlation between even reflections 0-18, as well as Tea fitness[2], and GDP per capita. Each reflection and the Tea indicator have been z-scored. As already mentioned, the

[2] A comment that we received to an earlier version of this report was that the (natural) log of Tea fitness may yield a better fit with GDP per capita, and possibly other economic variables. Therefore, the appendix provides alternative versions of the Figures in this section that use ln(Tea Fitness). These figures are rather heavily influenced by outliers, which is why we prefer the non-log versions in the main text.

reflections converge quickly, as the correlation between reflection 2 and 4 is already 0.987. The scatterplots show a clearly positive relationship between the country indicator and GDP per capita, at all reflections. Correlations vary between 0.532 at Reflection 0, and 0.678 at Reflection 8. For the Tea fitness indicator, the (linear) correlation is 0.531, but this appears to be a non-linear and heteroskedastic relation.

Figure 2.2 shows the relationship between HH reflection 8 and the Tea fitness indicator. This is a fairly tight relationship, although clearly nonlinear. The (linear) correlation is 0.758. With the exception of a few countries with very low Tea fitness core, the relationship between Tea and HH reflection 8 is exponential, i.e., if we would graph Tea fitness on a logscale, a more or less straight line would result. This suggests a very close similarity between the two approaches as far as country scores are concerned.

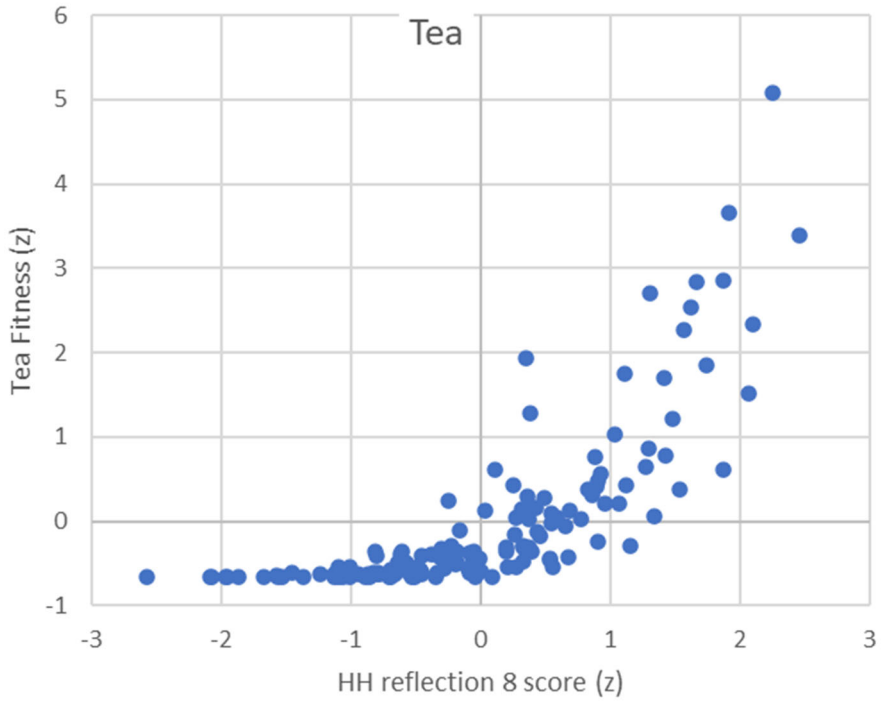


Figure 2.2. HH reflection 8 and Tea Fitness graphed against each other

Hidalgo and Hausman report regressions in which the dependent variable is a country's growth rate of GDP per capita, and in which the initial log-GDP per capita and the country complexity indicator (at various reflections) are the explanatory variables. We repeated those regressions and the basic results are documented in Table 2.1. We also include results with the Tea fitness indicator instead of the HH reflection. For short-period growth rates (3 years and 5 years), *t*-values for the country-complexity indicator (including Tea fitness) are generally low (not significant at 10% level) and declining with the reflection number. For longer growth rates, especially 20 years, we do see significant *t*-values, although this is never the case for the Tea indicator.

Table 2.1. *t*-values of explanatory variables in regressions explaining growth of GDP per capita

	3y growth rate		5y growth rate		10y growth rate		20y growth rate	
	logGDPpc	F	logGDPpc	F	logGDPpc	F	logGDPpc	F
Reflection 0	-0.84	1.29	-1.52	1.87*	-2.27**	1.69*	-4.63***	2.77***
Reflection 2	-0.34	0.34	-1.03	0.93	-1.68*	1.07	-4.61***	2.93***
Reflection 4	-0.24	0.22	-0.90	0.78	-1.55	0.99	-4.37***	2.67***
Reflection 6	-0.20	0.17	-0.84	0.72	-1.50	0.96	-4.19***	2.46**
Reflection 8	-0.18	0.15	-0.80	0.68	-1.48	0.93	-4.04***	2.27**
Tea Fitness	-0.75	1.45	-1.06	1.29	-1.58	0.66	-3.52***	1.35

Note: regressions include log of GDP per capita (logGDPpc) and Hidalgo-Hausman country complexity (F) at various reflections, and Tea Fitness as explanatory variables. HH reflections and Tea fitness are z-scored. Growth rates are 3 years (3y), 5 years (5y), 10 years (10y) and 20 years following initial period. Robust standard errors used to calculate *t*-statistics.

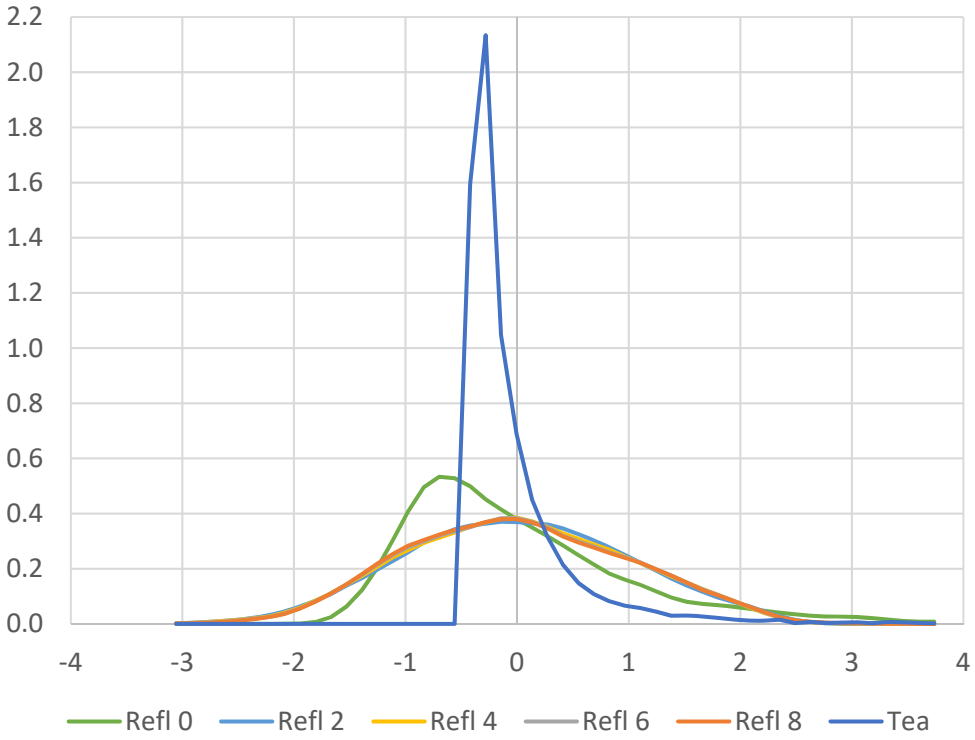


Figure 2.3. Kernel density estimation for the even reflections 0-8 of the product indicator in Hidalgo and Hausman and product Quality in Tea (all indicators z-scored)

Next, we look at the distribution of the product indicator at Reflections 0-8 and Tea quality. We present the smoothed histogram (by kernel density) of this indicator in Figure 2.3. The product indicator has been z-scored. The figure shows that as far as the HH reflections are concerned, the distribution is very similar for reflections 2 – 8, but slightly different for reflection 0. At reflection 0, the distribution is skewed to the left, but for higher reflections, it becomes rather symmetrical

around zero (although still slightly left-skewed). The Tea quality indicator is very peaked, at a negative (i.e., below-average) value. Compared to its own left tail (which is short due to the truncation of the non-z-scored indicator at zero), the right tail is fairly long and fat, but compared the HH reflections (especially 2-8), the right-tail of the Tea indicator is not particularly heavy.

While Figure 2.3 shows that the HH reflection indicators and the Tea product Quality indicator differ in terms of the distribution, these indicators also differ in terms of the values for individual products. This is shown in Figure 2.4, which graphs the HH reflection 8 on the horizontal axis vs. Tea product quality on the vertical axis. The difference is striking, as the correlation is slightly negative. All of the very high values of Tea quality are associated with below-average HH values.

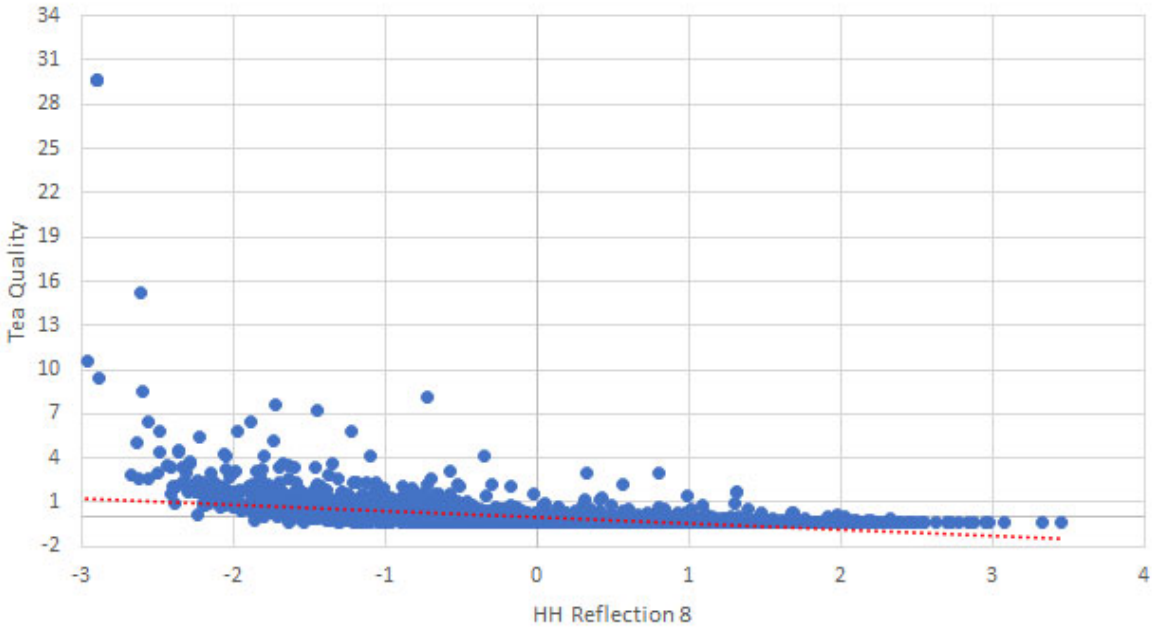


Figure 2.4. HH and Tea product level indicator graphed against each other (1998)

Finally, Figure 2.4 shows the intertemporal stability of the country- and product-indicators. It graphs the results of one year (1998) against those of the next (1999). We see high stability (a close relationship with high correlation) for both country-level indicators. However, at the product level, the stability of the indicators differs greatly between HH (reflection 8, but results are similar for other reflections) and the Tea indicator. The relationship for HH reflection 8 is less strong at the product level (R=0.932) than at the country level (R=0.991), but overall, there is a very clear correlation for both levels. For Tea, on the other hand, the country-level correlation is as high as HH, but at the product level the positive correlation mostly results from about a dozen large values. The correlation for the Tea product indicator is 0.600. This is the result of a large number of values that have a relatively high score in one year, but a low score in the next year, or vice versa. For example, of all product that had a positive z-scored Tea value in 1999 (28% of all products), 30% had a negative z-scored Tea value in 1998.

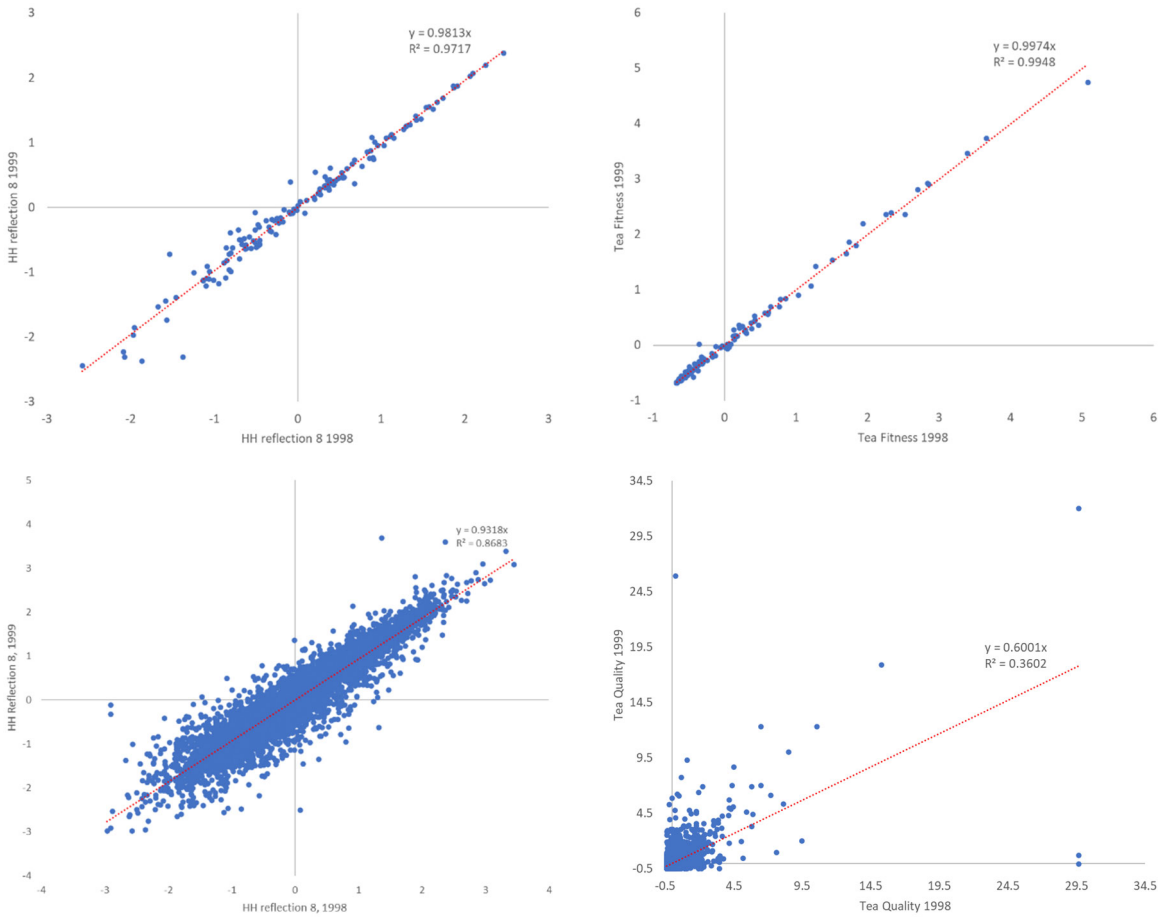


Figure 2.5. Temporal stability of country- and product-indicators, HH reflection 8 and Tea

2.2. A role for canonical correlation

Our motivation for introducing canonical correlation analysis in complexity analysis is twofold. On the one hand, we observe that both complexity measures that we discussed so far produce results that are highly variable over time, especially for products (as illustrated above). While part of the time-variability may be due to variability in the underlying data (we will return to this possibility in Section 5, where we use a smoothed RCA indicator), it is also clear that the HH and Tea methods tend to amplify this variability. The method we will propose below does not include such amplification, and hence is expected to produce results that are inherently more stable.

Our second reason for introducing canonical correlation analysis is related to the interpretation of the complexity indicators. It was shown that at the country level, the indicators derived from HH and Tea are correlated, and also that each of these indicators is correlated strongly to GDP per capita. However, they arrive at the country-level indicators on the basis of very different assessments of which products are relevant. This discrepancy between the product level scores of the two methods is all the more striking given that the broad ideas underlying the methods is similar. Neither method is strongly theory-based, but they both start from the idea that economic development essentially means developing new production capabilities, and that these capabilities reveal themselves in the data about how many and which kind of countries are able to produce a certain product with revealed comparative advantage. However, even with this common starting idea, the algorithms that are proposed by Tea and HH produce very different results at the detailed level of export products.

In terms of the language of machine learning, both the HH and Tea methods are examples of unsupervised learning, which means that their algorithms are attempts to find patterns in the trade data without using any information on the variables that represent the process of economic development, such as GDP per capita and its growth rate. Only after the algorithm has provided the indicator(s) based on the trade data is a correlation to economic development presented and interpreted. While we see no principal reasons to reject such an approach of unsupervised learning from the data, the comparison of HH and Tea results also seems to suggest that the specific nature of the algorithm has a large influence on the outcomes.

The approach that we develop in the remainder of this paper proposes that such algorithmic "confusion" may be avoided by adopting a rather simple algorithm based on the principle of supervised learning. In supervised learning, data on the development process is included from the very beginning, so that the basic proposition becomes to find data patterns in which development (however we will measure it) and trade specialization are correlated.

A supervised learning approach has the additional advantage that all dimensions in the data on economic development can be used next to each other. For example, we noted that while the correlation of the country-level indicators to GDP per capita is strong in both Tea and HH, growth of GDP per capita is not so universally strongly correlated to the country-level indicators. Thus, if the HH and Tea complexity measures are related to "competitiveness" of nations, then we may ask what competitiveness actually means. Does it mean a high living standard, or high growth rates of the living standard? And what if we wanted to include other dimensions of competitiveness, such as sustainability? In their basic form, and due to their nature of unsupervised learning, the Tea and HH algorithms do not provide any opportunity to include such considerations into the analysis.

In the Tea method, however, the results from the basis algorithm are proposed to be used in "a data-driven method—the selective predictability scheme—in which we adopt a strategy similar to the methods of analogues, firstly introduced by Lorenz, to assess future evolution of countries." (Cristelli et al, 2015, p. 1). This method establishes a link between the country level indicator from the algorithm and GDP per capita as well as its growth rate. This is an additional algorithm, which we consider to be outside the scope of our analysis in this paper. Our proposed supervised learning algorithm is a multi-dimensional alternative to the combination of basic Tea algorithm plus the selective predictability scheme.

3. Method

In this section, we will summarize the Canonical Correlation Complexity (CCC) method. Subsections will cover the basic algorithm and the use of the algorithm for predictions.

3.1. The basic algorithm

The basic idea of the CCC method is that we may specify competitiveness as an N-dimensional phenomenon, and hence for each country in the analysis obtain a vector of length N that measures competitiveness, and that this competitiveness dataset may be related to higher-dimensional dataset on product-level specialization. Product-level specialization is measured in the same way as in the HH and Tea methods, i.e., by the matrix **M** as defined in section 2 above. Typically, this

matrix has a number of products in the order of magnitude of 5,000, i.e., there are many more products than countries. We also assume that the number of dimensions of competitiveness (N) is much lower than either the number of countries or the number of products.

Canonical Correlation Analysis (CCA) is a method that produces weighted sums in both datasets (the matrix **M** and the matrix with data on country competitiveness), in such a way that the correlation coefficient between these weighted sums is maximized (see, e.g., Wilks, 2008). More specifically, we write the matrix **M** as cxn (i.e., countries in the rows and products in the columns), and the matrix **L** of size cxN contains the competitiveness dataset. In its most basic form, CCA then finds matrices **A** (NxN) and **B** (nxN), which we use to obtain matrices **LA** (cxN) and **MB** (also cxN). **A** and **B** are weight matrices, which are chosen in such a way that the correlation coefficient between the first columns of **LA** and **MB** is maximized, and given this relationship, the correlation between the second columns is maximized, etc.

A problem with this basic procedure is that generally $n \gg c$, which means that there are too many degrees of freedom in choosing the contents of the **B** matrix. This is why we enhance the method with one additional step, which comes at the very beginning of the algorithm. This step is to perform a Principal Component Analysis (PCA) of the **M** matrix, which reduces the dimension of the product space. In this way, we transform the **M** matrix to $\mathbf{M}^*$ which is cxn^*, where n^* can be chosen freely as a cut-off, and $n^* < c$. We then perform the CCA with matrix $\mathbf{M}^*$ instead of **M**. Note that each column of $\mathbf{M}^*$ is a different linear combination (i.e., weighted sum) of all columns of **M**, and these weights (provided as an outcome of the PCA procedure) ensure that the columns of $\mathbf{M}^*$ are orthogonal to each other. In formal terms, the algorithm is as follows:

1. Construct de-meaned (in the country dimension) versions of matrices **M** and **L**. We will continue to use these same symbols, but will understand these two basic matrices to be de-meaned.
2. Perform PCA on matrix M, and choose the number of retained components (n^*) such that a fraction f of the total variance is retained, or, if this is more binding, $n^* < c$. The value f is a parameter that we vary in our sensitivity analysis. The PCA also yields a matrix **V** (nxn^*) of product loadings (i.e., this matrix contains the weights used to obtain the product scores). Essentially, **V** consists of the eigenvectors of the matrix product $\mathbf{M^T M}$ that are associated with its largest n^* eigenvalues, as stacked horizontally in decreasing order of the associated eigenvalues.
3. Construct the matrix $\mathbf{M}^* = \mathbf{MV}$, which contains the scores of all countries on the retained components in the PCA.
4. Perform CCA on matrices **L** and $\mathbf{M}^*$, yielding the weight matrices **A** and **B**, as well as a vector **r** that contains N canonical correlation coefficients. The elements of r can be seen as a measure of goodness of fit for each dimension that is derived in the CCA.
5. Calculate the matrix **VB** (nxN), which contains the product complexity scores for each product, in each dimension produced by the CCA. These product complexity scores (of which we have more than a single) are comparable to the product-level indicators in the HH and Tea methods.
6. Calculate the matrix **MVB** (cxN), which contains the country-level complexity scores, again in each dimension of the CCA. Again, we have more than a single of these country-level indicators, and each one is comparable to the country-level indicators of HH and Tea.

While this description of the algorithm has been written with a single year of data in mind, it may also be readily applied to pooled (panel) data. In this case, in the basic matrices **M** and **L**, matrices for individual years are stacked left-to-right. The algorithm then runs unchanged on these pooled data matrices. This is the approach we will take in the empirical analysis below.

3.2. Predictions (and rotation)

Our approach will be to perform the algorithm as documented in the previous subsection for a period T0, and then use the results of this estimation to predict the values that make up the competitiveness matrix **L** for a period T1>T0. We choose T0 and T1 in such a way that we have actual data for T1, so that we can compare the (out-of-sample) prediction to the actual data in order to evaluate the quality of the predictions.

In predicting, we can choose to either predict the values of the composite factors (**LA**) generated by the CCA, or the underlying individual indicators that make up the matrix **L**. We opt for the latter, because this gives a more direct picture in the form of the indicators that we are interested in. In order to obtain this prediction at the indicator level, we need to perform what we call a rotation of the CCA results. The entire algorithm for prediction is as follows:

1. Run the algorithm as specified in the previous subsection for a dataset for period T0 (this can be a pooled sample of years).
2. Having used the matrix $\mathbf{M_0}$ (period T0, or "in-sample") for the estimation of the model coefficients, use matrix $\mathbf{M_1}$ (period T1), the estimated matrix product **VB** (estimated from T0 data), the estimated vector of canonical correlations (again from T0), and the estimated weight matrix **A** (again from T0) to calculate $\mathbf{\hat{L}_1 = M_1 VB\tilde{r}A^{-1}}$. $\mathbf{\hat{L}_1}$ is the matrix of predictions for the competitiveness indicators for period T1 (see Wilks, 2008), $\mathbf{\tilde{r}}$ is the matrix with the vector of canonical correlations on the main diagonal and zeros elsewhere, and $\mathbf{A^{-1}}$ is the inverse of matrix *A*. Note that, in the calculation of $\mathbf{\hat{L}_1}$, the matrix product $\mathbf{M_1 VB}$ is similar to the last step of the basic algorithm in the previous subsection, and $\mathbf{\tilde{r}A^{-1}}$ represents a rotation that is used to obtain indicator-level predictions (see, e.g., Glahn, 1968, for an explanation of how such regression-based predictions are equivalent to CCA).
3. Because the predictions $\mathbf{\hat{L}_1}$ are de-meaned, we need to add a vector of means to obtain data that can be compared to the actual data of T1. We use the country means of period T0 (or the last year of T0 is the data are pooled) for this purpose.
4. Optionally, we can choose to also calculate the in-sample prediction errors for each column of **L** (i.e., $\mathbf{e_0 = L_0 - \hat{L}_0 = L_0 - M_0 VB\tilde{r}A^{-1}}$). If there is persistence in terms of these errors (i.e., if countries systematically over- or underperform relative to the estimated values), adding these in-sample prediction errors may improve the quality of the out-of-sample prediction errors (i.e., then, the ultimate out of sample predictions become $\mathbf{M_1 VB\tilde{r}A^{-1} + e_0}$ + the means of the variables in the period-T0 competitiveness dataset $\mathbf{L_0}$).

4. Selecting the *f* threshold

In the results that we present in this section and all following sections, we will use a slightly different data source for the detailed trade data. We still draw our data from Tacchella (2020), but in this case we use one of the smoothed RCA indicators presented there. We opt for the forward-looking indicator derived from the hidden Markov model (RCAf). This indicator smooths out yearly variations in the data by assuming that there is an underlying state of the RCA indicator which can vary stochastically on an annual basis (the method estimates the underlying state). We also do not binarize the RCAf indicator, but instead use it in its original form, which allows variation between 0 and 1, with 0.5 as the "neutral" value (corresponding to 1 in the case of the original RCA definition).

The use of the non-binary RCAf indicator vs the use of tRCA (as in Section 2 above) does not matter much for the results from our own method. We switch to RCAf, however, because we noticed that alternative methods (HH reflections and the Tea method) are much influenced by the use of non-binary RCAf. They tend to produce much more stable outcomes (as compared to, for example, Figure 2.5) when non-binary RCAf is used, and hence the comparison of our own method to HH or Tea is more conservative if we use non-binary RCAf.

We use four indicators in the competitiveness dataset. One of these is GDP per capita (as defined before). A second is the growth rate of GDP per capita, which we calculate as a forward-looking variable, e.g., the growth rate assigned to the year 2015 is the average annual growth rate of GDP per capita over the period 2015 to 2015 + G, where G is a growth lag that can be varied (we will work mostly with $G = 3$, but the results do not change much for larger values of G. A third variable is CO_2 emissions in metric tons per capita. This refers to CO_2 emissions within the territorial borders of the country, i.e., it also includes production for exports, but not CO_2 related to imports. The fourth and final variable is the growth of per capita CO_2 emissions, defined in the same was (and with the same value for G) as in the case of GDP per capita.

In order to determine which value of the threshold f to use in our subsequent analysis, we have to consider two factors. On the one hand, a larger value of will tend to increase the predictive power (at least in-sample), because a larger f means that more data are used in the analysis. On the other hand, a larger value of f may decrease the stability of the product complexity scores over time, because for each individual year, these scores become more adapted (over-fitted) to the specific values of the variables in the competitiveness data set (i.e., matrices $\mathbf{M_0}$ and $\mathbf{L_0}$).

We will analyze this tradeoff by performing the analysis on our pooled dataset with varying f. We pool the data for the period 1996 – 2012. A period of G=3 years is used to calculate growth rates (of GDP per capita and of CO_2 per capita) in the competitiveness dataset, i.e., the last growth rates that are used are for the period 2012 – 2015. We use the results of these estimations to predict (out-of-sample) the competitiveness data for the year 2015, which includes growth rates for 2015 – 2018. As a measure of the predictive power of the analysis, we use the RMSE for the prediction for each of the 4 competitiveness variables.

As a measure for stability of the product complexity scores, we divide the sample in two parts (1996 – 2003 and 2005 – 2012, i.e., we skip the middle year 2004), and correlate the product complexity scores between these two analyses. This correlation coefficient is used as the measure for stability of the scores. The results are documented in Table 4.1. The first column documents f,

which ranges from 0.4 to 0.99 in steps of 0.05, except the last step which is 0.04 (we cannot perform the analysis with f = 1). The other columns of the table document the RMSE and correlation (corr) for each of the variables. We also present this information in four graphs, one for each variable in the competitiveness dataset, in Figure 4.1.

Table 4.1. Stability and predictive power as a function of f

	GDP pc		g GDP pc		CO2 pc		g CO2 pc	
f	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE
0.40	0.956	0.908	0.947	3.023	0.948	1.252	0.939	4.465
0.45	0.926	0.896	-0.127	3.035	0.868	1.238	0.898	4.550
0.50	0.863	0.886	0.397	3.068	0.870	1.237	0.297	4.533
0.55	0.833	0.782	0.335	3.059	0.798	1.057	-0.070	4.502
0.60	0.827	0.676	0.381	3.087	0.809	0.941	-0.113	4.500
0.65	0.820	0.652	0.307	3.094	0.801	0.915	0.021	4.550
0.70	0.809	0.637	0.253	3.093	0.810	0.885	0.029	4.540
0.75	0.752	0.616	0.191	3.137	0.758	0.854	-0.067	4.638
0.80	0.769	0.569	0.145	3.172	0.759	0.791	-0.040	4.688
0.85	0.707	0.500	0.165	3.211	0.729	0.699	-0.043	4.855
0.90	0.621	0.463	0.234	3.254	0.646	0.656	-0.045	4.993
0.95	0.579	0.463	0.181	3.254	0.604	0.656	-0.038	4.993
0.99	0.579	0.463	0.181	3.254	0.604	0.656	-0.038	4.993

Notes: The column label "Corr" denotes the correlation coefficient between product complexity scores (our measure for stability), the column label "RMSE" denotes the root-mean-squared-error of the out-of-sample prediction for the variable (our measure for predictive power).

These results confirm the tradeoff between stability and predictive power for the two level-variables (GDP per capita and CO_2 per capita). For these two variables, increasing the value of f generally decreases stability (the "Corr" columns in the table tend to decrease) and also increases predictive power (the "RMSE" columns decrease). In the graph we see an upward-sloping relationship, with the small dots appearing on the righthand side. Thus, for these two variables, large (small) f values tend to have lower (higher) predictive power but higher (lower) stability of the product complexity scores.

The tradeoff does not exist for the two growth variables. These results are erratic, and if anything, there is a negative relationship between the RMSE and "Corr" values for varying f values (this is seen in the graphs as a downward-sloping relationship). This outcome is related to the general difficulty in predicting the growth rates of the variables in the competitiveness data set, a phenomenon that will be documented below.

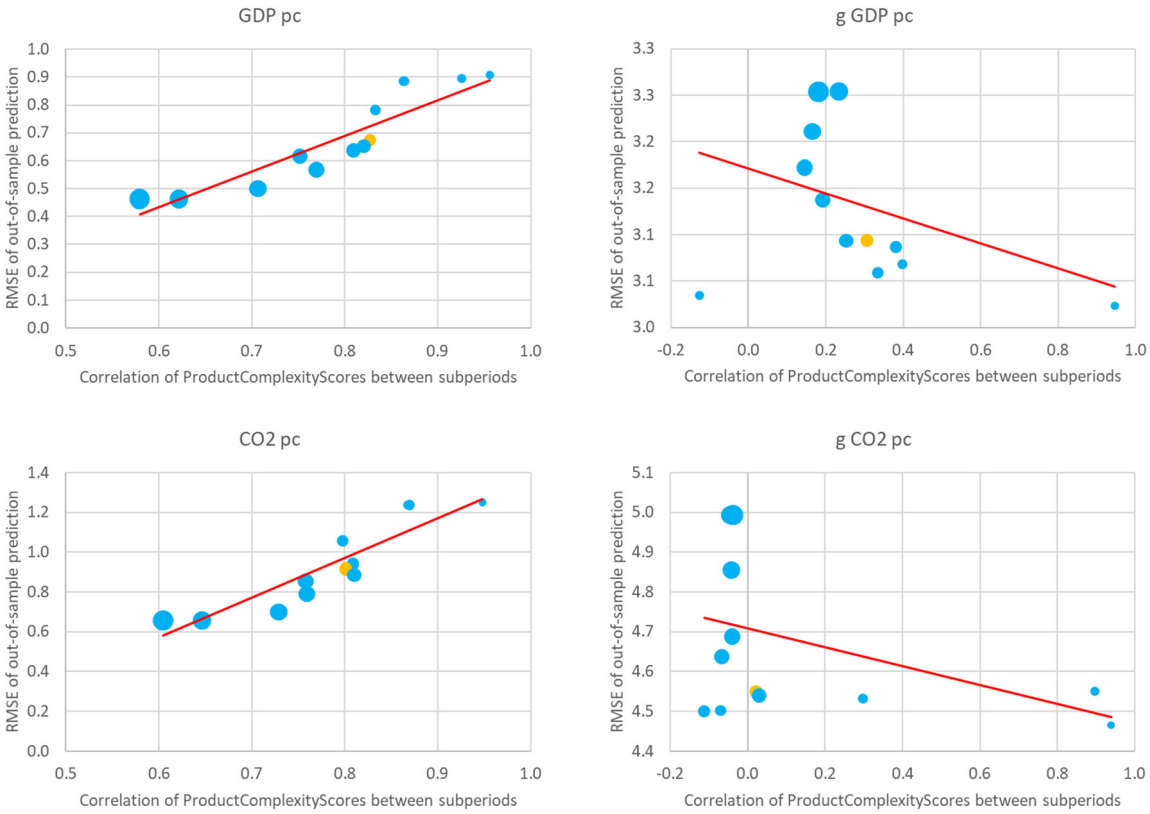


Figure 4.1. Stability and predictive power as a function of *f*

Note: the size of the dots represents the value of *f* (smaller/larger dots represent smaller/larger values)

Looking at these results, we opt for $f = 0.65$, which is an intermediate value. The product complexity score stability value ("Corr") for $f = 0.65$ is rather high (slightly above 0.8), and it is still fairly high (0.3) for the growth rate of GDP per capita. For the growth rate of CO_2 per capita, it is very close to zero. However, the RMSE values for the two growth rates are still below the middle of the range in Table 4.1, while they are more or less at the middle of the range for the two 'level' variables. Other values than $f = 0.65$ sacrifice either stability or predictability in a fairly drastic way.

5. CCC results: a basic summary

In this section, we present the basic results of the CCC estimation on the pooled sample as explained in the previous section, with $f = 0.65$. In Table 5.1, we report the loading matrix *A*. This matrix shows how the four competitiveness variables are weighted to produce new dimensions that correlate maximally with similar dimensions in the high-dimensional trade dataset. The first of these canonical correlations is high (0.83), and puts almost exclusive weight on the GDP per capita variable (this is the first column of the table; all other weights than the first are very close to zero). This confirms the impression we also obtained using the Tea and HH methods in Section 2, which is that the structure of (revealed) competitive advantage is very strongly related to the general development level of a country.

Table 5.1. The loading matrix (*A*) for the competitiveness dataset, and canonical correlations

	Factor 1	Factor 2	Factor 3	Factor 4
GDP per capita	0.827	-1.894	0.890	0.090
Growth of GDP per capita	0.012	0.037	0.298	-0.031
CO2 per capita	-0.022	1.505	-0.664	0.057
Growth of CO2 per capita	-0.008	0.010	-0.042	0.106
Canonical correlation	0.830	0.642	0.257	0.129

The second correlation has a strongly negative weight for GDP per capita, a strongly positive weight for CO_2 per capita, and weights very close to zero for the two growth rates variables. This particular structure is the result of the fact that these new dimensions of the data are orthogonal, and the fact that GDP and CO_2 per capita are strongly correlated to each other. These correlation coefficients of the raw data are documented in Table 5.2. Thus, while the first dimension captures the majority of the variance in GDP per capita and CO_2 per capita by simply putting a large weight on GDP per capita, the second dimension captures the remaining variance by weighting CO_2 per capita strongly, and GDP per capita negatively. This correlation is much lower than the first, but still sizable.

Table 5.2. The correlation matrix for the competitiveness dataset

	GDP per capita	Growth of GDP per capita	CO2 per capita	Growth of CO2 per capita
GDP per capita	1			
Growth of GDP per capita	-0.134	1		
CO2 per capita	0.925	-0.047	1	
Growth of CO2 per capita	-0.145	0.405	-0.189	1

The third correlation, which is again much lower than the second, has reverse signs on the CO_2 and GDP per capita variables, as compared to the second dimension), but weight the growth rate of GDP capita much higher. All in all, this dimension is a bit difficult to interpret (high development level, moderate on growth, and low on CO_2 per capita). The same holds for the last correlation, which is rather low, and has low loadings on all four variables (growth of CO_2 per capita has the largest loading).

All in all, these loadings are not particularly attractive in terms of providing an interpretation of common factors that summarize the development and growth experience of countries in an easy way. The first dimension ("development level") does so, but this is at a very basic level, and subsequent dimensions are difficult to interpret. This is partly due to the correlation structure of the four competitiveness variables (Table 5.2), in which the two 'level' variables are strongly correlated, the two growth rates are also correlated but weaker than the levels, and the correlations between levels and growth rates are negative, but very low in absolute value.

Given this finding, it is attractive to look at the rotated loading matrix, $\mathbf{rA}^{-1}$, which, when pre-multiplied with the product complexity scores, produces new dimensions that are maximally correlated to each of the four original pure variables, rather than weighted dimensions. It should be stressed that this rotation doesn't change any of the basic results of the CCA procedure, it is only a different way to represent these results. Because of the link to the pure variables, interpreting the CCA results in terms of these rotated loadings is always straightforward. Table 5.3 presents these results. While the numbers in this table do not have any obvious interpretation, we document them for completeness, as they will be the basis for the predictions that will be documented in the next section.

Table 5.3. The rotated loading matrix ($\mathbf{rA}^{-1}$) for the competitiveness dataset

	Factors (labeled by variable)			
Variable	GDP per capita	Growth of GDP per capita	CO2 per capita	Growth of CO2 per capita
GDP per capita	1.027	-0.380	1.313	-1.691
Growth of GDP per capita	-0.002	0.906	0.401	0.056
CO2 per capita	-0.011	0.852	-0.036	0.282
Growth of CO2 per capita	0.010	0.146	0.001	1.249

Next, Table 5.4 shows the correlations between the rotated product complexity scores. These are key result variables from the analysis, because they are the predictors (based on the trade data) for the competitiveness variables. The table shows that the procedure is very well able to reproduce the correlations between the level variables GDP and CO2 per capita (compare Table 5.2 to 5.4). The correlation between the two growth rates is a little lower (in absolute value) in Table 5.4 than in 5.2. On the other hand, the correlations between one growth rate variable and one level variable are somewhat higher (again in absolute value) in Table 5.4 than in Table 5.2.

Table 5.4. The correlation matrix for the product complexity scores, based on rotated loadings

	GDP per capita	Growth of GDP per capita	CO2 per capita	Growth of CO2 per capita
GDP per capita	1			
Growth of GDP per capita	-0.207	1		
CO2 per capita	0.921	0.064	1	
Growth of CO2 per capita	-0.482	0.311	-0.450	1

We now turn to the issue of stability of the product complexity scores between the two sub-periods, as already analyzed in Section 4 (Figure 4.1) for varying levels of *f*. Here we present more detailed results for $f = 0.65$, which are documented in Figure 5.1. As could be expected from the results in the previous section, we see that the correlation between the product complexity scores for the two sub-periods is very tight for both level indicators (GDP per capita and CO_2 emissions per capita), but not very strong for the two corresponding growth rates. For the two 'level' variables, it is interesting to see that the extreme values (low or high) share in the high correlations. This means that products that score very high (or very low) in one period have corresponding scores in the other period. This is less the case for the two growth rate variables.

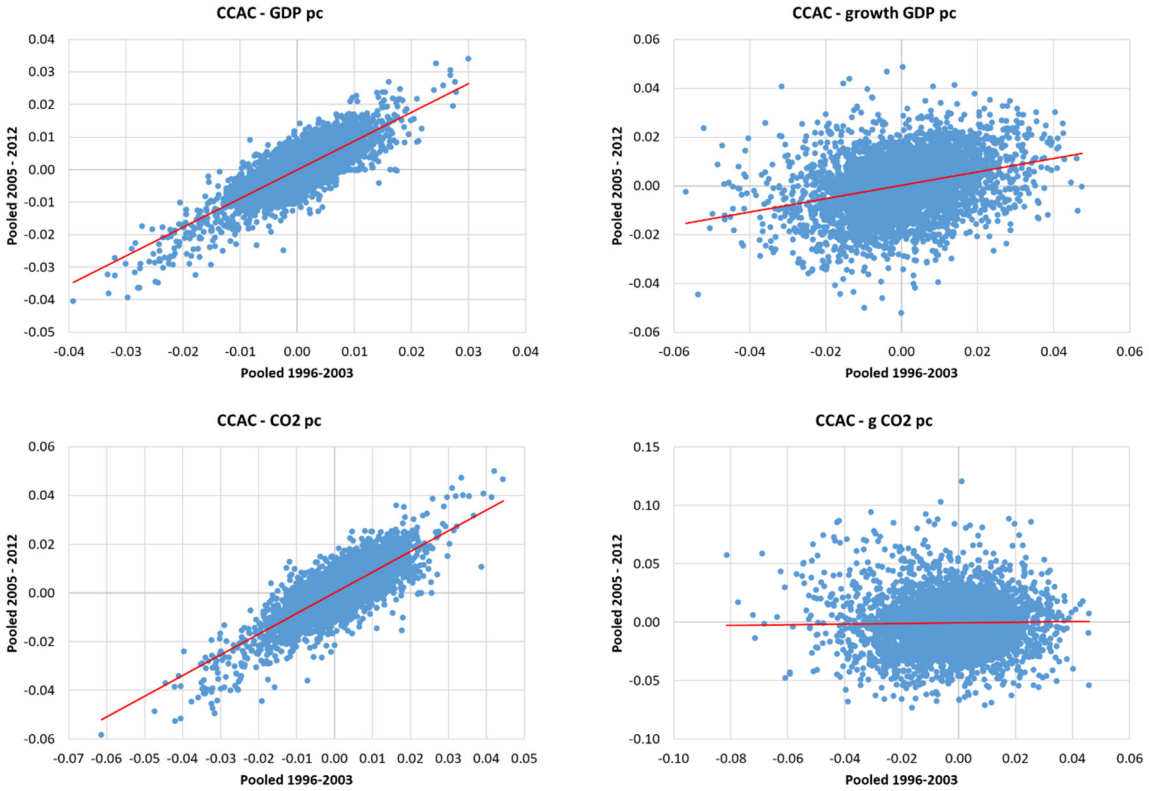


Figure 5.1. Stability of the four rotated product complexity score vectors (CCC)

Figure 5.2. shows the same stability plot for the Tea and HH (reflection 8) measures. Here we see a slightly different picture as compared to Section 2. The results in Figure 5.2 show much more stability as compared to Section 2. This is due to the use of a different indicator for the trade data (RCA) between the two sections. Here we use the RCAf indicator from Tacchella (2020), which is a smoothed indicator, while in Section 2 we used an RCA indicator based on the raw (non-smoothed) trade data. Thus, it is clear that the use of the smoothed RCA indicator improves stability for the HH and Tea measures very substantially. Nevertheless, in both indicators, there is a sizable subset of products which clearly fall outside the main relationship between the two periods. Interestingly, these are, in both cases, products that score relatively low in the early period, and high in the late period (the reverse doesn't happen except for a handful of cases for the Tea indicator).

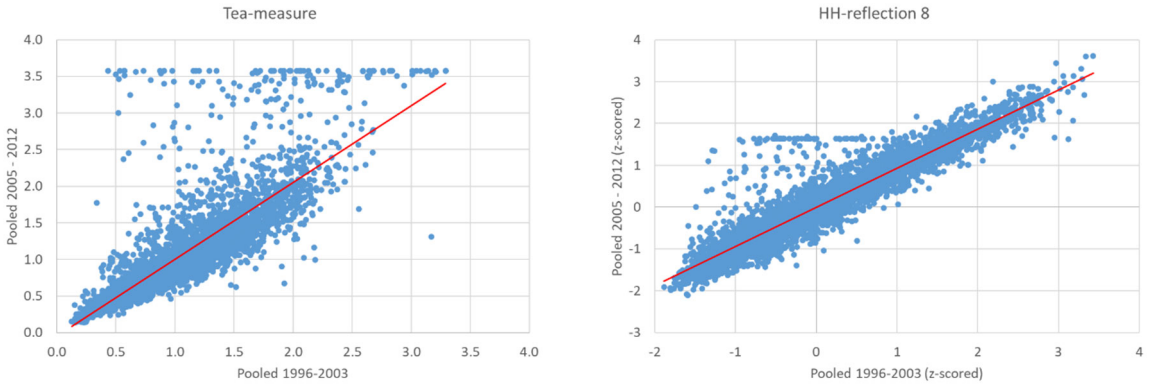


Figure 5.2. Stability of the Tea and HH (reflection 8) product indicator

We conclude this section by comparing the product level indicators to each other. Figure 5.3 compares the four CCA Complexity measures against each other. In the top-left figure, we see that there is a negative association (although with significant variance) between the product complexity scores for GDP per capita and those for the growth rate of GDP per capita. This is a crucial finding, because it points out that export products that are associated with high living standards tend to be associated with low growth rates. Or, phrased in the reverse way, export products that tend to be found in rapidly growing countries are not the same as those found in highly developed countries. This finding is fully compatible with the technology gap theory of growth (see, e.g., Fagerberg and Verspagen, 2021), and associated theoretical frameworks such a product life cycle theory of trade (e.g., Vernon, 1966). These theoretical approaches to economic growth stress the notion of structural change, which is the tendency that countries tend to change their structure of production (and consumption) while they are developing. Development takes place by gradual adoption (and adaptation) of foreign knowledge, leading to gradual structural change, rather than by direct imitation of the products that highly developed countries tend to export.

In the top-right figure, we see that the product complexity scores for GDP per capita are strongly correlated to those for CO_2 per capita. This means that by and large, exports products that are associated with high living standards are also associated with a high CO_2 footprint, something which is, of course, not surprising. In the middle-left figure, the product complexity scores for GDP per capita are negatively correlated to those for the growth of CO_2 emissions. Given the previous two figures, this is exactly as expected. The middle-right picture shows a flat line representing the relationship between the product complexity scores for the growth rate of GDP per capita and those for the level of CO_2 per capita. This suggests that the product that are associated to high growth are not related in any way to CO_2 intensity.

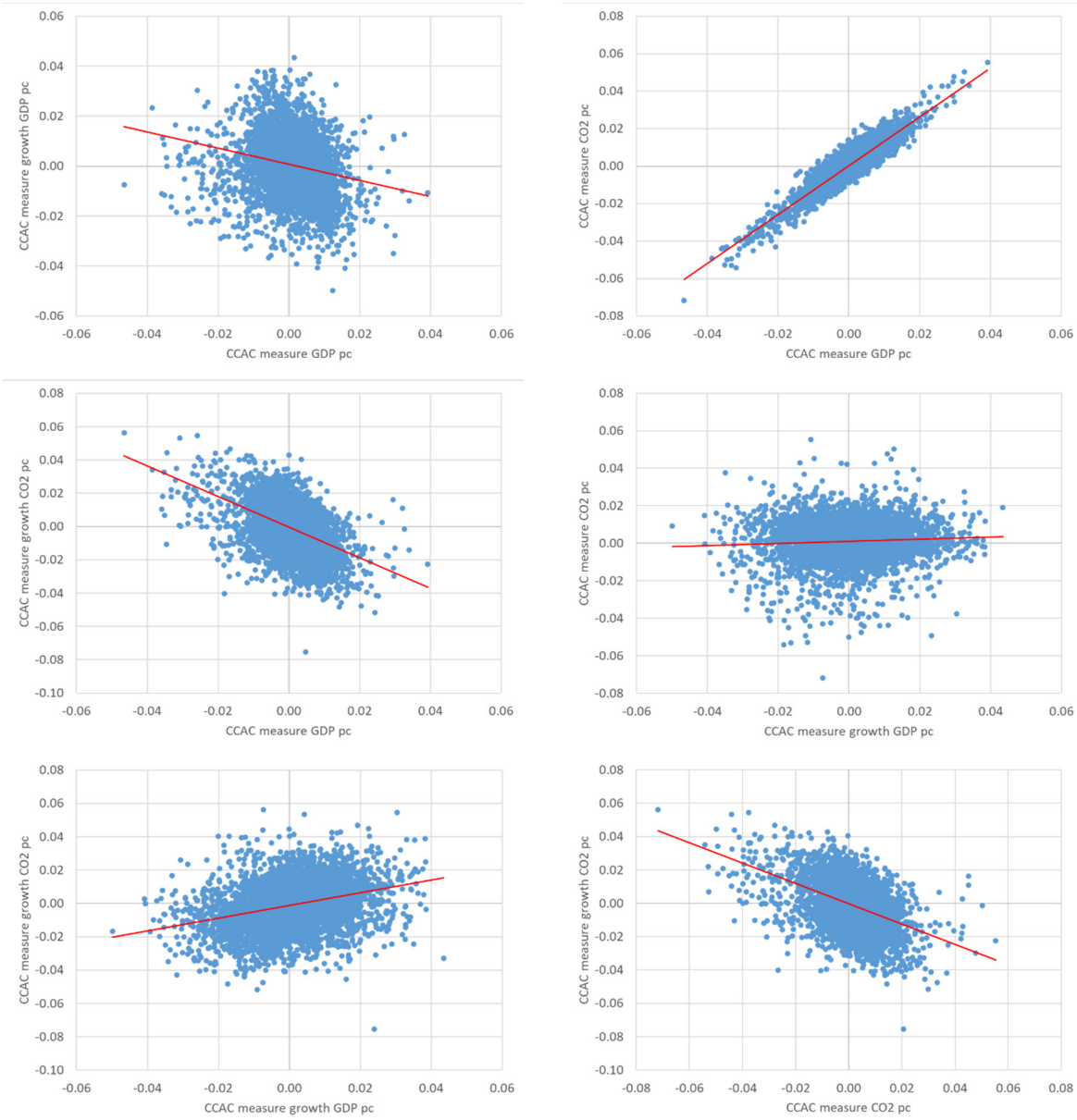


Figure 5.3. Comparing the four CCC product complexity scores to each other

On the other hand, in the left-bottom picture, the product complexity scores for growth of GDP per capita and growth of CO_2 emissions, are weakly and positively correlated, suggesting that products that are associated to growth also tend to be associated to growth of emissions (but this correlation is weak). Finally, in the bottom-right picture, the product complexity scores for CO_2 per capita are related negatively to those for the growth rate of that same variable.

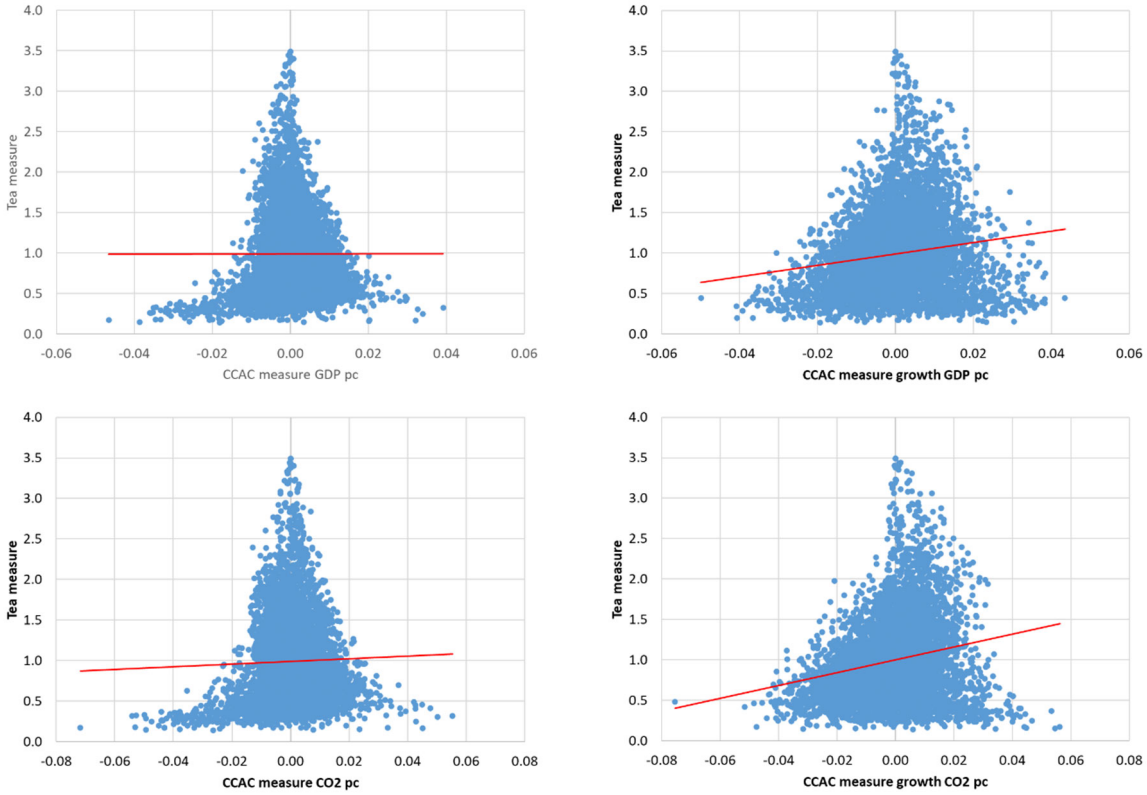


Figure 5.4. Comparing the CCC product complexity scores to the Tea product measure

In Figure 5.4, the four product complexity scores variables are compared to the (single) Tea product indicator. All of these show a tent-shaped cloud of observations, although the shape of the tent differs between the four cases. Such a tent shape is peculiar, because it suggests that the products that score high on the Tea indicator tend to have intermediate values for each of the four CCC indicators (but note that the tent is "filled", i.e., not all of the intermediate value products in the CCC measures have high Tea values). Clearly, the overall conclusion is that the Tea indicator and the CCC indicators point to very different conclusions with regard to the question the nature of export products and their relationship to growth and development.

Figure 5.5 shows the relationship between the HH product level indicator (reflection 8 as before) and the four product complexity scores for CCC. These graphs show more variety than the ones for the Tea measure, with some pointing to weakly positive relationships and others to weakly negative relationships, or a flat relationship. Overall, like in the case of the Tea indicator, the CCA complexity product complexity scores point to very different conclusions than the HH indicator.

Finally Figure 5.6 shows the relationship between the HH product level indicator and the Tea indicator. This is similar to Figure 2.5, although we now use a pooled dataset and a different indicator for RCA. Despite this, the relationship remains weak and negative, although we now see a number of outliers above the tight cloud of observations, suggesting perhaps a tilted U shape. In any case, this graph confirms the overall conclusion that none of the three available product complexity indicators point in the same direction.

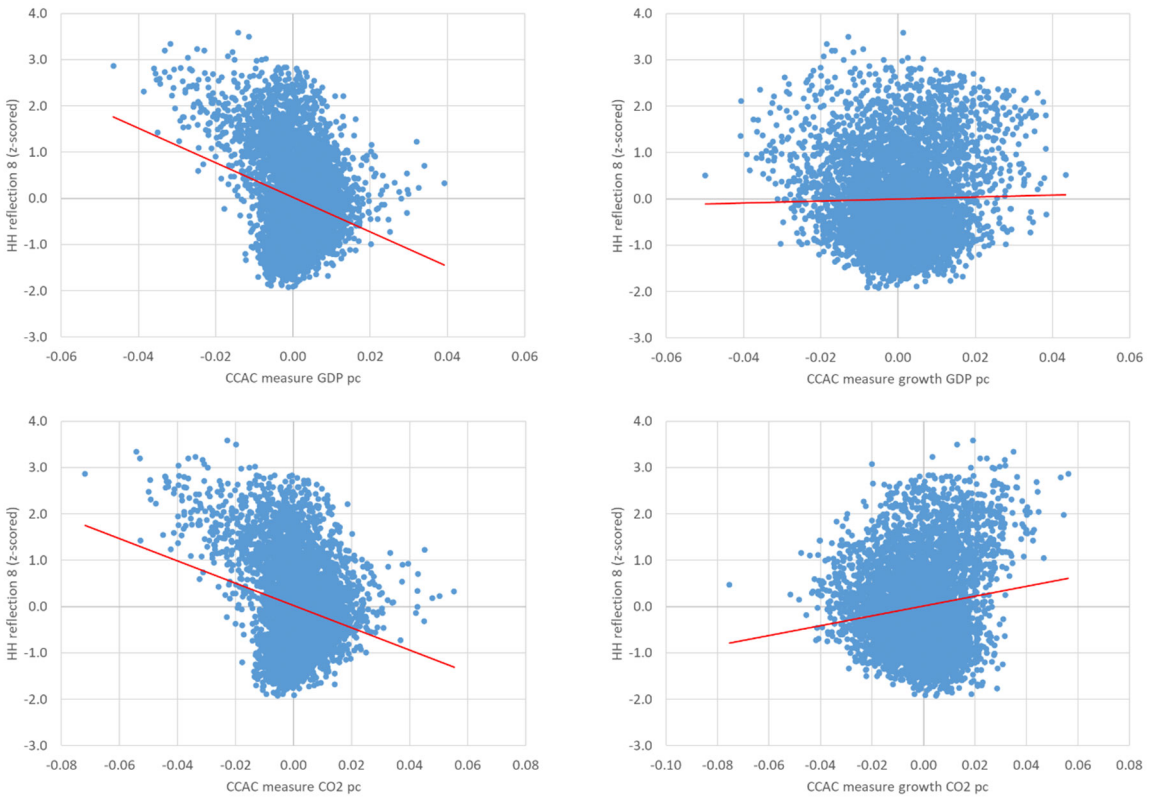


Figure 5.5. Comparing the CCC product complexity scores to the HH reflection 8 product measure

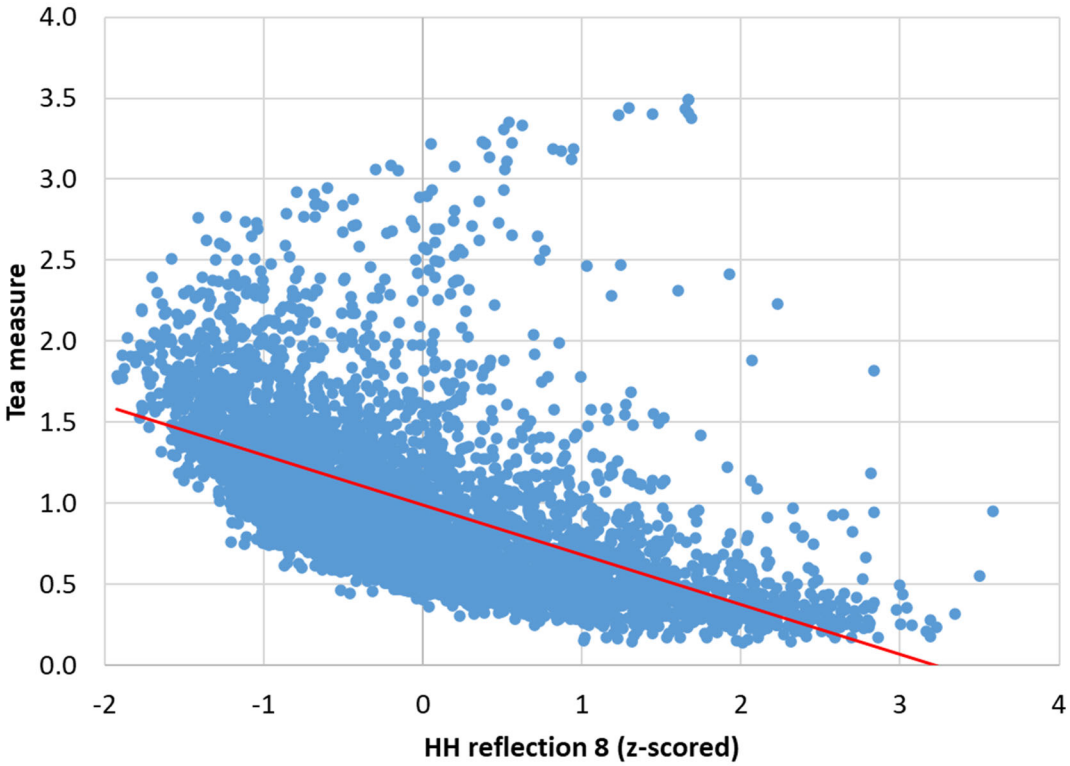


Figure 5.6. Comparing the Tea product measure to the HH reflection 8 product measure

6. Predictive power

We will now present results for the predictive power of the CCC method. Prediction will be aimed the variables in the competitiveness data set, i.e., GDP per capita and CO_2 per capita, and their respective growth rates. We will also provide results for the predictive power of the Tea method and the HH method, so that a comparative view will emerge.

We will take the rotated country scores for the competitiveness variables as the predictions, as explained in Section 3. These country scores have zero mean, because we de-mean the variables before they go into the CCA procedure. For in-sample prediction, which is where we focus first, we add the means to the predictions, so that we can compare to the actual values of the competitiveness variables. These results are presented in Figure 6.1, which, for each of the four variables puts the predicted value on the horizontal axis, and the actual value on the vertical axis. These graphs use the entire pooled dataset, i.e., for every country there are 17-yearly observations.

The R^2 statistics for the fitted regression lines are provided in the graphs, and can be taken as a rough indication of the comparative quality of the prediction. In this respect, the quality of the prediction for GDP per capita seems highest. Here the fit is relatively tight, although there are particular areas where some of the predictions seem to be off. This the case for a group of observations between values 7 and 9 on the horizontal axis, where the actual value is much lower than the predicted value, and for a group of observations between 8 and 11 on the horizontal axis, where the actual value is higher than the predictions. The graph for CO_2 per capita is very similar to GDP per capita, as could be expected by the high correlation between those two variables.

However, the predictions for the growth rates are not nearly as good as those for the two 'level' variables. Here, the R^2 values are much lower (0.16 is the highest value). However, here are a number of outliers (negative as well as positive) in the real data that the predictions do not follow, and these influence the R^2 somewhat. Nevertheless, the quality of the growth rate predictions is not very good, with, for example, a high number of cases where the actual growth rate is negative, but a positive rate if predicted, or vice versa.

In Figure 6.2, we turn to out-of-sample predictions. Here we use the product complexity scores from the pooled sample for 1996 – 2012, and then predict the competitiveness variables by using the 2015 RCA data, as well as the estimated (and rotated) loadings (the procedure is described in full in Section 3). Because this produces de-meaned competitiveness variables, we also need to estimate the mean of the variables in 2015. We do this by estimating an autoregressive model on the pooled data and then using this model to forecast the 2015 values (This is done using Matlab's 'ar' and 'forecast' functions, the lag used in the model is 2 years). Because this simply adds a fixed number to the de-meaned predictions for each country (for each variable), this procedure does not influence the predictions much beyond a cosmetic way.

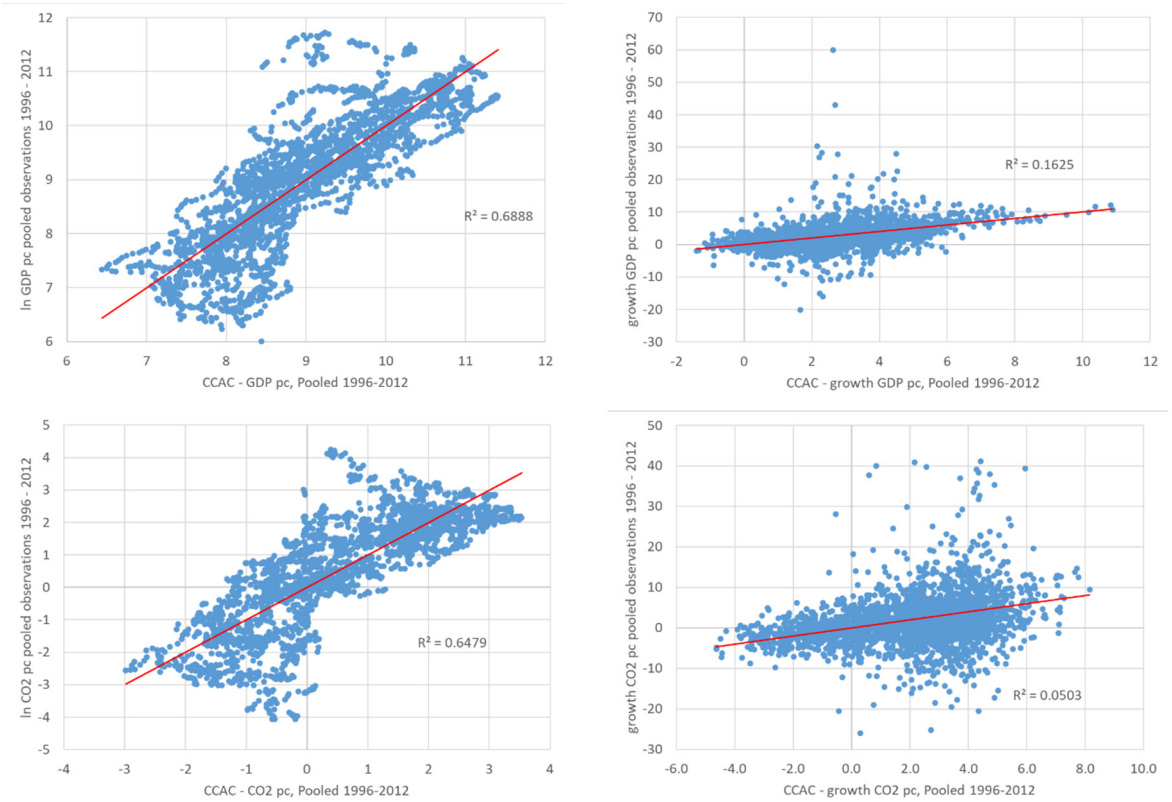


Figure 6.1. In-sample prediction of competitiveness variables by CCC method

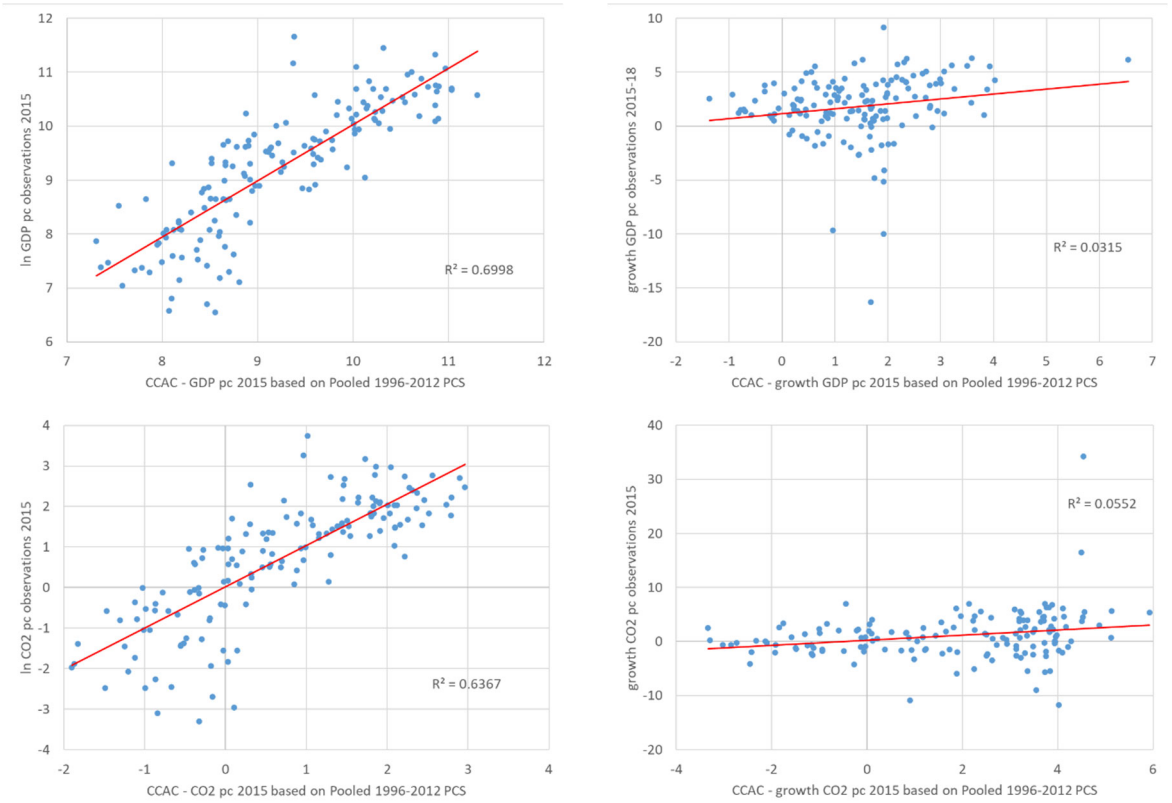


Figure 6.2. Out-of-sample prediction of competitiveness variables by CCC method

Figure 6.2 shows that the out-of-sample predictions for the two 'level' variables (GDP per capita and CO_2 per capita) are quite good as compared to the in-sample predictions. The R^2 values of the regression lines that are fitted in the graphs are comparable to those in Figure 6.1 (but note the difference in the number of observations, as we now have only a single observation per country). The out-of-sample predictions for the two growth rate variables are much weaker, however. In this case, the correlations between the predicted values and the actual values are rather low. We must therefore conclude that the CCC method can predict the level competitiveness variables relatively well, but not so the growth rate variables.

In Figure 6.3, we present a slightly different form of the out-of-sample predictions. Here we use what we call an in-sample residual, and add it to the out-of-sample prediction of Figure 6.2. The general idea of this procedure is that there may be, at the country-level, a systematic and persistent error in the CCC predictions. Such a systematic and persistent error may then be reflected in the residuals, i.e., the difference between the actual value of an observation and the in-sample prediction for that observation. We use a similar autoregressive model as we used to predict the sample mean of the competitiveness variables to predict the in-sample residual for each country and each variable. This predicted in-sample residual is then added to the out-of-sample prediction.

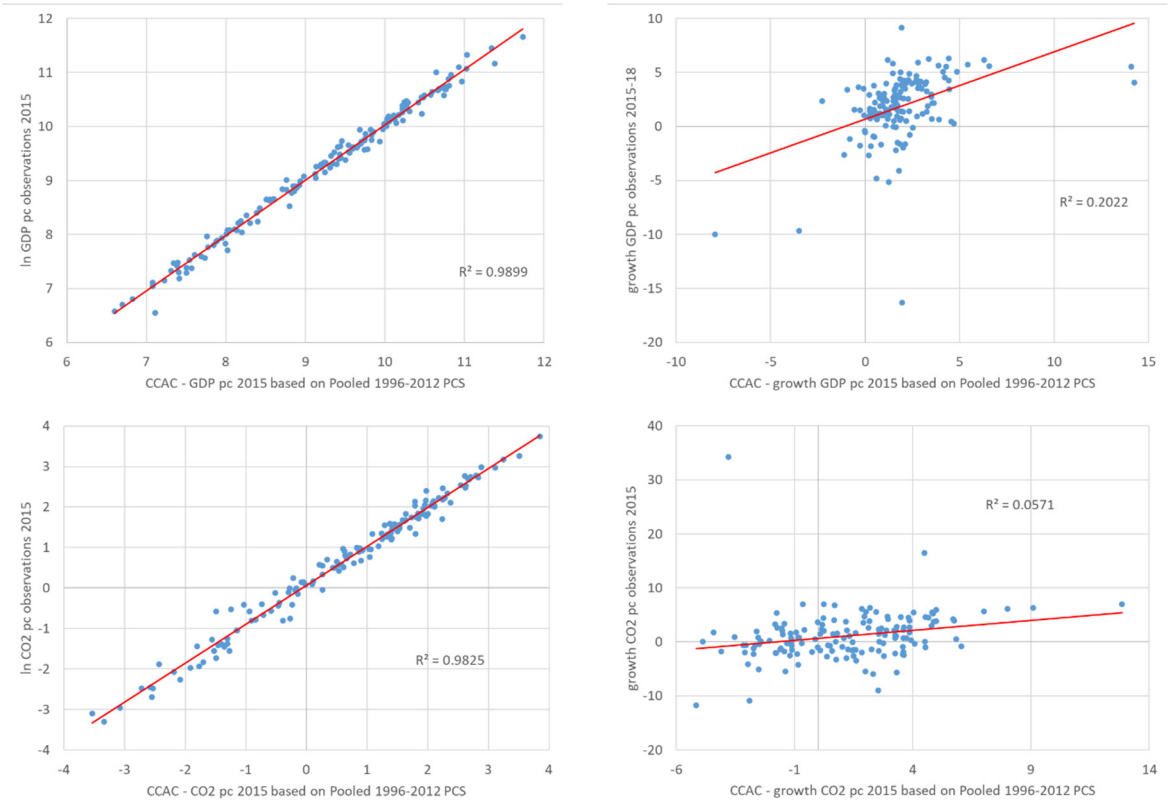


Figure 6.3. Out-of-sample prediction of competitiveness variables by CCC method, with in-sample residuals

Figure 6.3 shows that this adds to the quality of the prediction. The two 'level' variables (GDP per capita and CO^2 per capita) now show very high correlations (in both cases $R^2>0.98$) between

predicted and actual values. Also, the quality of the prediction of the growth rate of GDP per capita is much increased, although the R^2 in this graph remains well below the value for the two 'level' variables. Only the growth rate for CO2 per capita remains very hard to predict even with the in-sample residuals.

We will also compare the predictions of the CCC method to the Tea and HH predictions. Note that both these methods have only one set of product-level scores (comparable to the product complexity scores of the CCC method). Therefore, we use the same prediction variable for all four competitiveness variables. This puts the Tea and HH methods at a disadvantage as compared to the CCC method, but is an unavoidable consequence of how the methods work.

In case of the Tea method, our predictions work as follows. We calculate the product complexity values on the basis of the pooled sample, and then apply these in combination with the 2015 RCA values to calculate country Fitness for 2015. This is then graphed against each of the four competitiveness variables, in Figure 6.4. Again, we report R^2 values in the graphs. Like in the case of the CCC method, predictions using Tea are much better for the two 'level' variables GDP per capita and CO_2 per capita than for the growth rates of these variables. Thus, compared to CCC, Tea does not help much in predicting growth, which was the weak point of CCC.

If we compare the R^2 values in Figure 6.4 to those in Figure 6.2, we see that CCC has clearly higher values (representing a higher quality prediction) than Tea for the two 'level' variables. For the two growth rates, the R^2 values are much more comparable, although Tea shows a marginally higher value for the growth rate of GDP per capita.

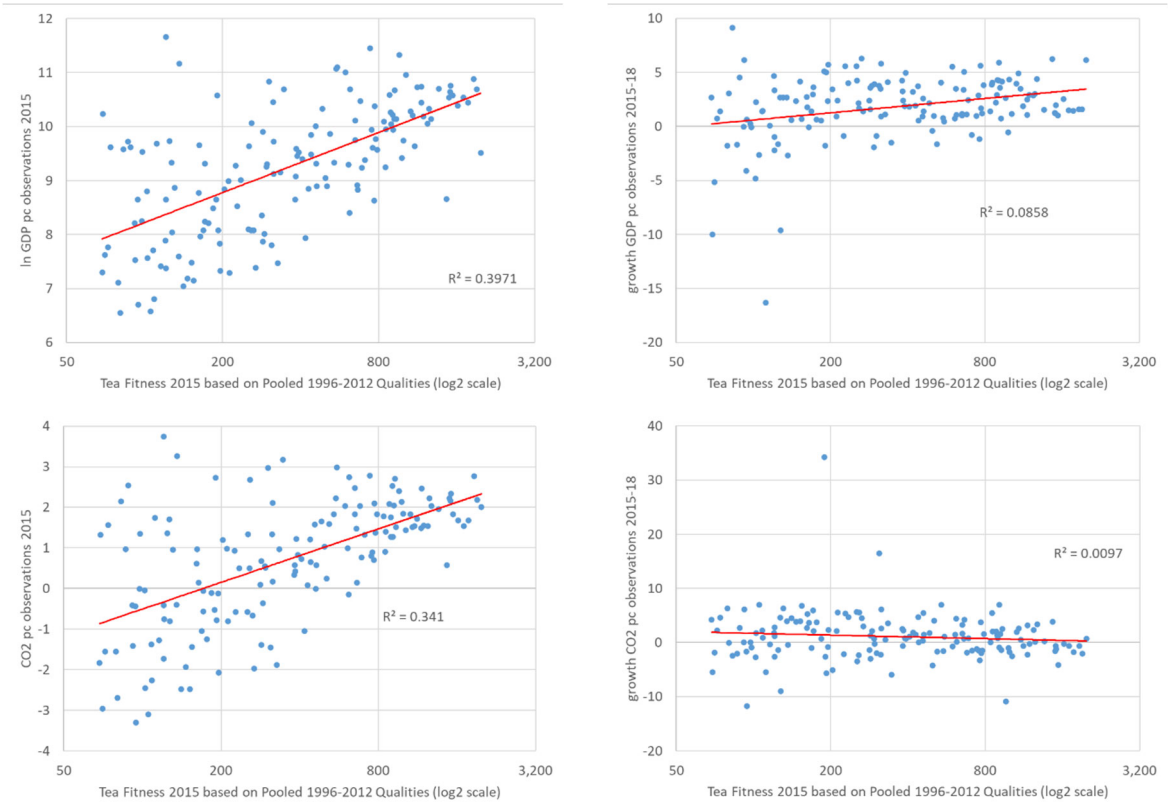


Figure 6.4. Out-of-sample prediction of competitiveness variables by Tea method

In Figure 6.5, we present predictions from the HH method. Here we use reflection 7 product scores from the pooled (1996 – 2012) sample in combination with 2015 RCA values to calculate country scores, which are then comparable to reflection 8 scores (note that by nature of the method, we need an even reflection for the country scores, which requires the preceding and hence uneven reflection for the product scores, see Section 2). Again, this is a single variable, instead of the 4 sets of country scores in CCC, and we plot this single indicator against all four competitiveness variables.

Like all predictions considered so far, the HH predictions are much better for the level variables GDP per capita and CO_2 per capita than for the growth rates of these variables. This is reflected in the R^2 values for the two 'level' variables, which are higher than the corresponding values for Tea, but lower than for CCC. The HH predictions for the growth rates are very weak, with the prediction for the growth rate of GDP per capita showing essentially a zero correlation to the actual data, and the growth rate of CO_2 per capita showing a mildly negative correlation.

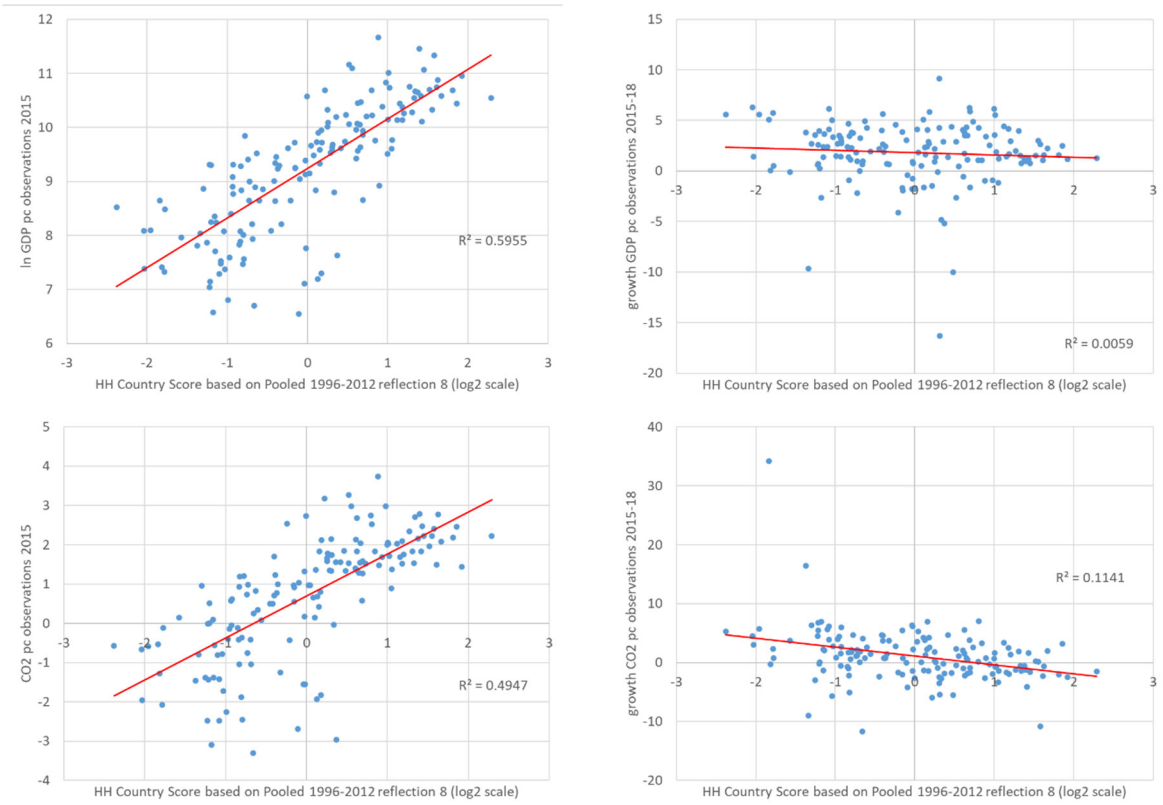


Figure 6.5. Out-of-sample prediction of competitiveness variables by HH method (reflection 8)

Overall, we conclude that the CCC method can predict the level of GDP per capita and CO_2 emissions per capita relatively well, but the growth rates of these variables much less so. If we use a similar prediction technique as we use for CCC to Tea and HH, we obtain better results using CCC as far as the level variables are concerned, and comparably weak results for the growth rates.

7. Some additional reflections on diversification and complexity

In this final section we delve a little deeper into the interpretation of the widely differing results that we obtained in the CCC method, the HH method and the Tea method with regard to the product scores (Figures 5.3 – 5.6). Each of these methods makes an attempt to score products in terms of the "complexity" or sophistication, and then weights these product scores together into a measure of country "fitness" (or simply "country scores"), using the country's diversification profile (i.e., its scores on the RCA indicators) as the weights. Thus, the country scores (fitness) consist of two basic underlying dimensions: the country's diversification profile (in how many products is it specialized?), and the product complexity scores of the products that it is specialized in.

Our aim in this section is to relate these two aspects of the country measures of each of the methods (our own CCC indicators, HH and Tea) to the (in-sample) predictions of these methods as presented in the previous section. In order to do this, we will first break down each of the country indicators into a part that is related to diversification, and a part that is unrelated to diversification. The non-diversification related part will, for example, capture the sophistication of products in terms of the value added to be captured in their markets. Then, we will use these two parts in regression models that have the variables from our country competitiveness data set (GDP per capita, CO_2 emissions per capita, and their growth rates) as the dependent variables. In these regressions, we will look at the effect size of each of the two parts of the country indicators (related to diversification and unrelated to diversification). We use the pooled sample (1996 – 2012) for this purpose.

In the first step of this approach, we look at the simple univariate regression of each of the country indicators. These correlations are shown in Figure 7.1, while the regressions are shown in Table 7.1. Remember that these regressions are only aimed to split the variance of the country indicators into a part that is related to diversification, and a part unrelated to diversification. We use the predicted values of these regressions as a representation of the part related to diversification, and the residuals for the part unrelated to diversification.

In these regressions, the Tea indicator stands out, as it has a very tight fit (high R^2), which indicates that for the Tea indicator, the part related to diversification is relatively high, and the diversification-unrelated part is very small. For the CCC prediction of the growth of GDP per capita, the opposite is true: here the fit is very low, indicating that the part (un)related to diversification is low (high). For the other four indicators (HH and the other three CCA indicators), the R^2 values are intermediate, indicating that the parts related and unrelated to diversification are both substantial.

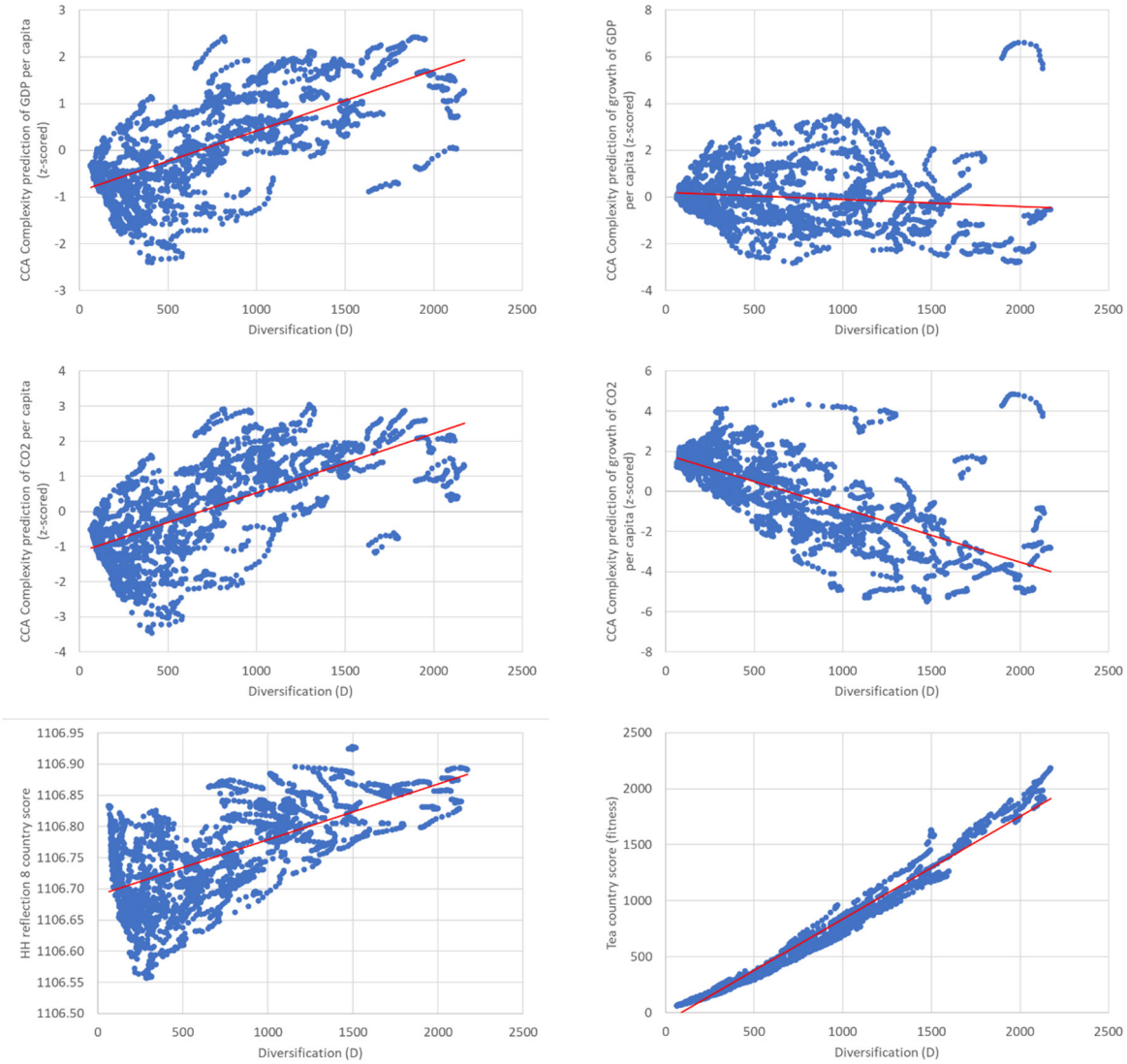


Figure 7.1. Diversification and the country score indicators (CCC, HH and Tea)

Table 7.1. Regressions models for the country score indicators (CCC, HH and Tea)

	Dependent variable					
	CCC GDP pc	**CCC gr GDP pc**	**CCC CO_2 pc**	**CCC gr CO_2 pc**	**HH**	**Tea**
D	0.001295 ***	-0.000301 ***	0.001688 ***	-0.002683 ***	0.0000891 ***	0.9166 ***
Cons	-0.8814 ***	0.2049 ***	-1.1488 ***	1.8260 ***	1106.69 ***	-78.28 ***
Adj. R^2	0.461	0.015	0.438	0.463	0.391	0.978

Note: three stars denote significance at the 1% level of better in a *t*-test.

In Table 7.2, we use the diversity- and non-diversity related parts of the various country score indicators as explanatory variables in regressions with the competitiveness variables as dependents. We do not document the full regression results, but instead only the (adjusted) R^2 and an indication of the effect sizes of the diversity and non-diversity related parts. For the effect size, we use the epsilon squared (ε^2) statistic (Kline, 2013), which measures the fraction of the variance of the "dependent" variable that can be attributed to variation in the explanatory variables (in our case, the diversity and non-diversity related parts of the country indicators). Note that the ε^2 statistics related to individual variables are measures for partial relationships, which means that the sum of $\varepsilon^2 - D$ and $\varepsilon^2 -$ non-D may be larger than $\varepsilon^2 -$ total. Note also that ε^2-total is a comparable measure to adjusted R^2, which is why the results f0or these two statistics are identical for the three decimals that are reported in the table.

Table 7.2. Size effects of (non-)diversity related parts of country score indicators (CCC, HH and Tea) in explaining competitiveness variables

	Country indicator					
	CCC GDP pc	**CCC gr GDP pc**	**CCC CO_2 pc**	**CCC gr CO_2 pc**	**HH**	**Tea**
			GDP per capita			
Adj. R^2	0.685				0.554	0.328
ε^2 – total	0.685				0.554	0.328
$\varepsilon^2 - D$	0.501				0.415	0.320
ε^2 – non-D	0.540				0.348	0.017
			Growth of GDP per capita			
Adj. R^2		0.118			0.002	0.002
ε^2 – total		0.118			0.002	0.002
$\varepsilon^2 - D$		0.002			0.001	0.001
ε^2 – non-D		0.116			0.001	0.001
			CO_2 per capita			
Adj. R^2			0.647		0.505	0.287
ε^2 – total			0.647		0.505	0.287
$\varepsilon^2 - D$			0.446		0.364	0.285
ε^2 – non-D			0.508		0.309	0.006
				Growth of CO_2 per capita		
Adj. R^2				0.045	0.024	0.021
ε^2 – total				0.045	0.024	0.021
$\varepsilon^2 - D$				0.021	0.021	-0.000
ε^2 – non-D				0.024	0.003	0.021

For the CCC indicators, we have one for each competitiveness variable (these are the in-sample predictors, as in Figure 6.1), while we relate the single HH and Tea indicators to all four competitiveness variables. For each of the competitiveness variables, the CCC indicator always has the highest adjusted R^2 (ε^2), which is due to the fact that, in the CCA procedure, this method

chooses the weights to maximize this correlation. For GDP per capita as the dependent variable, the CCC and HH indicators both show relatively balanced ε^2 statistics for diversity and non-diversity. For CCC, the non-diversity effect size is slightly larger than the diversity effect size, while for HH this is the other way around. For the Tea method, on the other hand, the effect size for the non-diversity part is very small. Obviously, this is mainly the result of the fact that the non-diversity part of the Tea indicator is very small.

For the growth of GDP per capita, the explained variance (R^2) is small, as was already seen above. Only the CCC method has some explanatory power here, and this is almost exclusively related to the non-diversity part of this indicator. For CO^2 per capita, the results are very similar to GDP per capita: CCC and HH have the highest explanatory power and both diversity-related and non-diversity-related parts play a relatively large role in this, while the Tea indicator shows almost no effect of the non-diversity related part. Similarly, the results for the growth of CO^2 emissions per capita are similar as for the growth of GDP per capita.

In summary, these results suggest that, in terms of their relation to diversification, the CCA complexity indicators and the HH indicator are relatively similar to each other, while the Tea indicator stands out. The CCC indicators and the HH indicator capture both diversification, and factors that are not related to diversification in the narrow sense. This seems to increase their ability to generate higher correlations to the variables in the competitiveness data set. On the other hand, the country-variation of the Tea indicator seems to be primarily related to variation in diversification levels of countries, which limits its ability to yield high correlations with the competitiveness variables.

8. Summary and conclusions

The CCC method is able to predict the level variables in our competitiveness data set, GDP per capita and CO_2 emissions per capita, rather well, both in sample and out of sample. The growth rates of these variables are not well predicted. Compared to other, pre-existing methods (Hidalgo and Hausman, 2009; Tacchella et al, 2012), the quality of the predictions by the CCC method, as measured by correlations to the actual data, is higher, also for the growth rate variables. This implies that the growth rate predictions of the alternative methods are not very strong either. Growth remains very hard to predict for the CCC method as well as other methods.

We also note that the in-sample residuals of the predictions have a fairly high degree of persistence. This means that these residuals are valuable additions to the out of sample predictions of the CCC analysis. Using the in-sample residuals in the out of sample predictions increases the quality of these out of sample predictions. This also holds for the predictions of the growth rate variables, although these remain much harder to predict than the level variables.

We also find that the product level scores, which are the main predictive tools of all three methods, differ widely between the methods. Thus, while all three methods have some predictive power for the level of GDP per capita (and CO_2 emissions), the micro details of the predictions differ widely. This makes it difficult to derive specific policy conclusions about which products are related to development, unless the policymaker is willing to put faith in just one of the proposed methods.

Even though our own method, CCC, produces the highest correlations to the actual data, our main policy advise would be to remain cautious with these complexity-based indicators.

In the interpretation of the differences in product level indicators between the methods, we have already hinted at the fact that none of the complexity methods is well rooted in the economic literature on development and growth. The methods, including our own, are data-driven rather than driven by theoretical insights about economic growth and its relation to structural change. A better integration between methodology and (economic) theory may well be the most fruitful way to proceed.

Nevertheless, the CCC method provides ample insight into which export products are associated with high or low levels of development, and the associated CO_2 emissions levels. These associations, or correlations, are stronger than for any of the two other methods. We therefore conclude that in terms of the basic question of how export structure and economic development are related, the CCC level provides the best available answer.

In terms of further research (also beyond the scope of the second report of the project), it will also be interesting to apply the selective predictability scheme method (Cristelli et al., 2015), which has been developed in the context of the method by Tacchella et al. (2012), but which can also be applied to the indicators of the CCC method.

References

Abramovitz, M. (1989). Thinking about Growth: And Other Essays on Economic Growth and Welfare (Studies in Economic History and Policy: USA in the Twentieth Century). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511664656.

Albeaik, S., Kaltenberg, M., Alsaleh, M. and C.A. Hidalgo, 2017. Improving the Economic Complexity Index, Papers 1707.05826, arXiv.org, revised Jul 2017

Albeaik, S., Kaltenberg, M., Alsaleh, M. and C.A. Hidalgo, 2017a. 729 new measures of economic complexity (Addendum to Improving the Economic Complexity Index), Papers 1708.04107, arXiv.org, revised Aug 2017

Cantore, N. and Alcorta, L., 2021, Structuralists' Contribution to Development Thinking: Old, New and Unresolved Issues, in: Foster-McGregor, N., Alcorta, L., Szirmai, A. & B. Verspagen (eds.), New Perspectives on Structural Change. Causes and Consequences of Structural Change in the Global Economy, Oxford: Oxford University Press.

Cristelli, M., Tacchella, A. and L. Pietronero, 2015, The Heterogeneous Dynamics of Economic Complexity. PLoS ONE 10(2): e0117174. doi:10.1371/journal.pone.0117174

Fagerberg, J. and B. Verspagen, 2021, Technological Revolutions, Structural Change and Catching Up, in: Foster-McGregor, N., Alcorta, L., Szirmai, A. & B. Verspagen (eds.), New Perspectives on Structural Change. Causes and Consequences of Structural Change in the Global Economy, Oxford: Oxford University Press.

Fagerberg, J., Srholec, M. & B. Verspagen, 2010, Innovation and Economic Development, in: Hall, B. H. and Rosenberg, N. (eds), Handbook of the Economics of Innovation, Elsevier, Amsterdam, pp. 833 - 872

Freire, C., 2021, Economic Complexity Perspectives on Structural Change, in: Foster-McGregor, N., Alcorta, L., Szirmai, A. & B. Verspagen (eds.), New Perspectives on Structural Change. Causes and Consequences of Structural Change in the Global Economy, Oxford: Oxford University Press (forthcoming 2021).

Glahn, H.E. 1968. Canonical Correlation and its Relationship to Discriminant Analysis and Multiple regression, Journal of the Atmospheric Sciences, vol. 25, pp. 23-31.

Hidalgo, C.A and R. Hausmann, 2009. The building blocks of economic complexity, Proceedings of the National Academy of Sciences, vol. 106, pp. 10570–10575.

Kline, R. B., 2013. Beyond Significance Testing: Statistics Reform in the Behavioral Sciences. 2nd ed. Washington, DC: American Psychological Association.

Nomaler, Ö. and B. Verspagen, 2021, Patterns of Diversification and Specialization in International Trade, in: Foster-McGregor, N., Alcorta, L., Szirmai, A. & B. Verspagen (eds.), New Perspectives on Structural Change. Causes and Consequences of Structural Change in the Global Economy, Oxford: Oxford University Press (forthcoming 2021).

Tacchella, A., 2020. Hidden Markov Models for RCA, European Commission, Joint Research Centre (JRC), Seville, Spain, mimeo.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & L. Pietronero, 2012, A New Metrics for Countries' Fitness and Products' Complexity, Scientific Reports, vol. 2, article number 723.

Vernon, R., 1966. International Trade in the Product Cycle. Quarterly Journal of Economics 80, pp. 190-207.

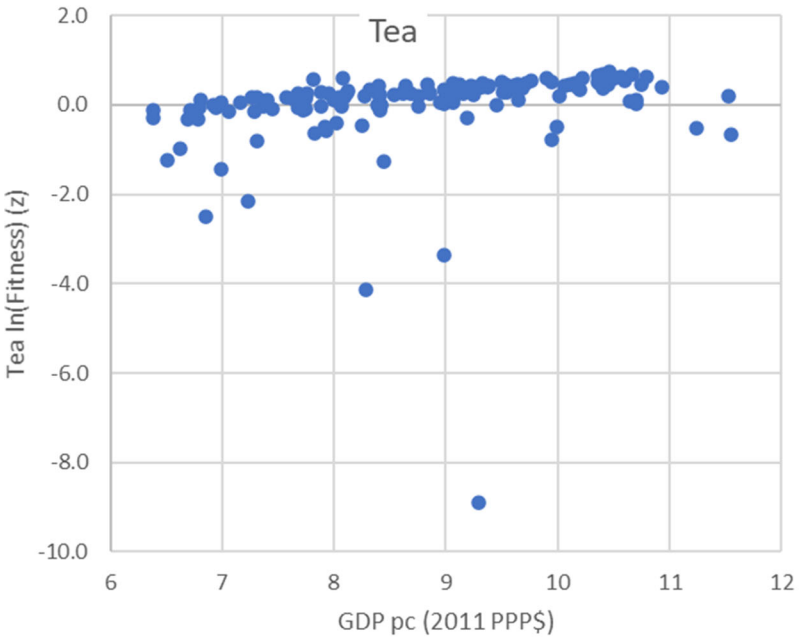
Verspagen, B., 2005, Innovation and Economic Growth, in: Fagerberg, J., Mowery, D.C. and Nelson, R.R. (eds), The Oxford Handbook of Innovation, Oxford: Oxford University Press, pp. 487-513.

Wilks, D.S., 2008. Improved statistical seasonal forecasts using extended training data, Int. J. Climatol. 28: 1589–1598

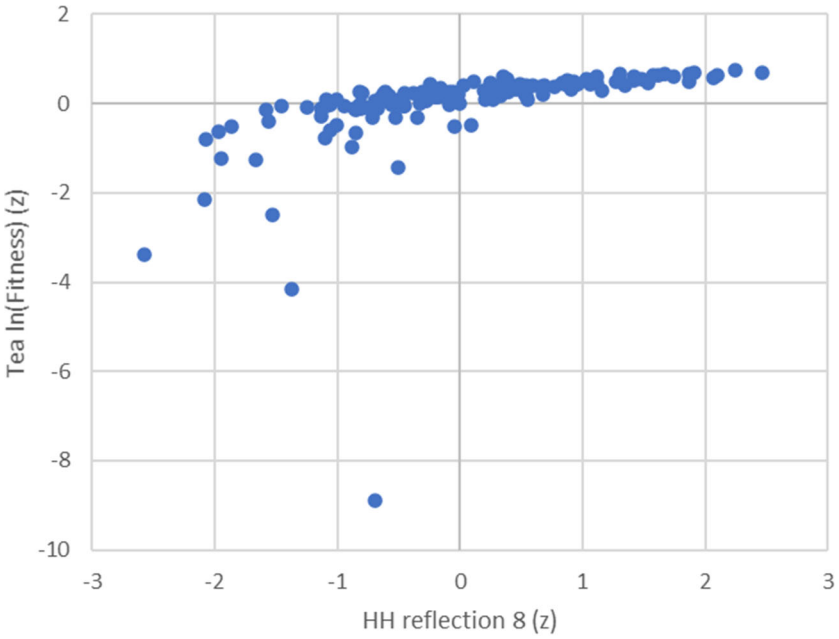
Appendix. Alternative figures from the main text with ln(Tea fitness) instead of tea fitness

This appendix contains alternative versions of all figures in the main text where Tea fitness appears. These alternative versions use the natural log (ln) of Tea fitness, instead of just fitness, which is used in the main text. Figures 6.4 and 6.5 in the main text already use ln of Tea fitness, and hence are not represented in this appendix.

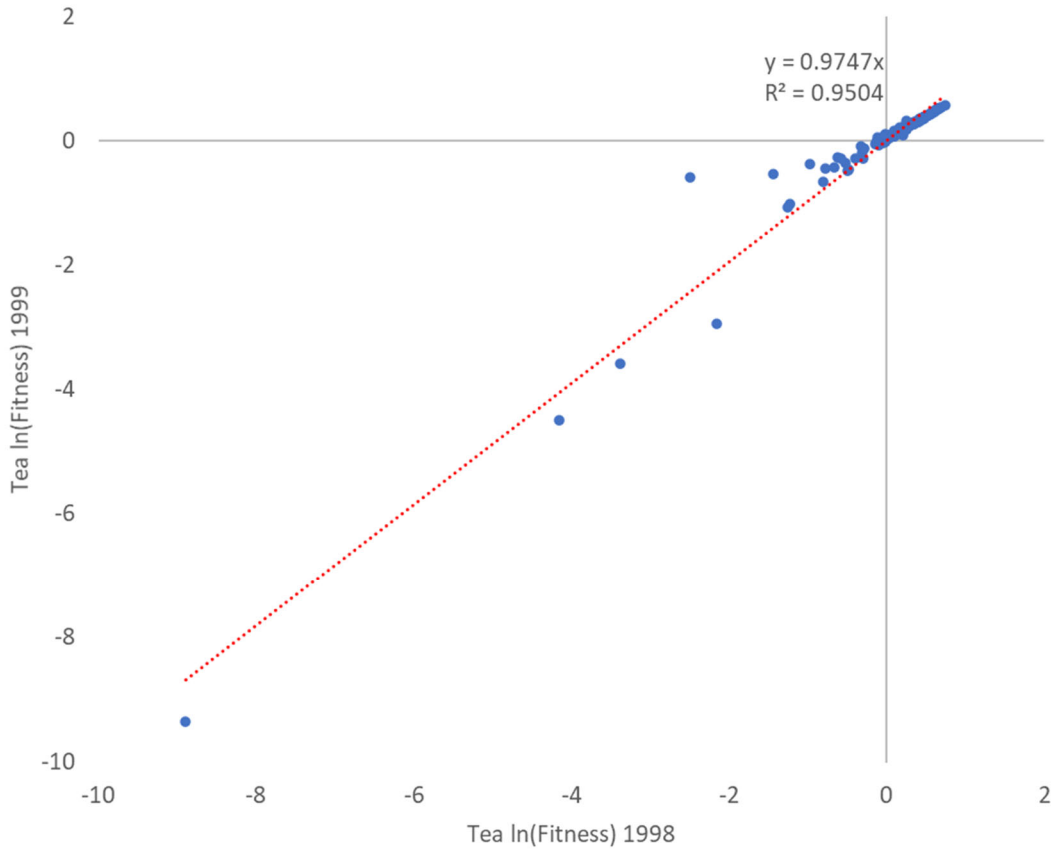
Below is the alternative version of the right-bottom panel of Figure 2.1 ("Even HH reflections 0-8 and Tea Fitness graphed against GDP per capita"):



Below is the alternative version of Figure 2.2 ("HH reflection 8 and Tea Fitness graphed against each other"):



Below is the alternative version of the right-top panel of Figure 2.5 ("Temporal stability of country- and product-indicators, HH reflection 8 and Tea"):



The UNU-MERIT WORKING Paper Series

2022-01 *Structural transformations and cumulative causation towards an evolutionary micro-foundation of the Kaldorian growth model* by André Lorentz, Tommaso Ciarli, Maria Savona and Marco Valente
2022-02 *Estimation of a production function with domestic and foreign capital stock* by Thomas Ziesemer
2022-03 *Automation and related technologies: A mapping of the new knowledge* base by Enrico Santarelli, Jacopo Staccioli and Marco Vivarelli
2022-04 *The old-age pension household replacement rate in Belgium* by Alessio J.G. Brown and Anne-Lore Fraikin
2022-05 *Globalisation increased trust in northern and western Europe between 2002 and 2018* by Loesje Verhoeven and Jo Ritzen
2022-06 *Globalisation and financialisation in the Netherlands, 1995 – 2020* by Joan Muysken and Huub Meijers
2022-07 *Import penetration and manufacturing employment: Evidence from Africa* by Solomon Owusu, Gideon Ndubuisi and Emmanuel B. Mensah
2022-08 *Advanced digital technologies and industrial resilience during the COVID-19 pandemic: A firm-level perspective* by Elisa Calza Alejandro Lavopa and Ligia Zagato
2022-09 *The reckoning of sexual violence and corruption: A gendered study of sextortion in migration to South Africa* by Ashleigh Bicker Caarten, Loes van Heugten and Ortrun Merkle
2022-10 *The productive role of social policy* by Omar Rodríguez Torres
2022-11 *Some new views on product space and related diversification* by Önder Nomaler and Bart Verspagen
2022-12 *The multidimensional impacts of the Conditional Cash Transfer program Juntos in Peru* by Ricardo Morel and Liz Girón
2022-13 *Semi-endogenous growth in a non-Walrasian DSEM for Brazil: Estimation and simulation of changes in foreign income, human capital, R&D, and terms of trade* by Thomas H.W.Ziesemer
2022-14 *Routine-biased technological change and employee outcomes after mass layoffs: Evidence from Brazil* by Antonio Martins-Neto, Xavier Cirera and Alex Coad
2022-15 *The canonical correlation complexity method* by Önder Nomaler & Bart Verspagen