UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics

Final Dissertation

Problems of Ranking and Dynamics of Complex Bipartite Networks

Thesis supervisor

Prof. Fulvio Baldovin

Thesis co-supervisors

Prof. Attilio L. Stella

Dr. Gianluca Teza

Candidate

Fabrizio Boninsegna

Academic Year 2021/2022

Introduction

The nowadays availability of large datasets and the recent increase in computational power offer a new paradigm to understand complex systems. However, data provide abundant detail that generally carry no labels on the procedure for extracting the important information that interests us about the system. Statistical physics and information theory offer a framework on which it is possible to manipulate large datasets on many different systems, leading to an interaction with other disciplines and creating a whole new branch called Complex System Science. This proficient union brought new insights in various fields such as ecology [1], population dynamics [2][3], ecosystem [4], nervous system [5] and economics, which is the framework in which we will move in this thesis.
A recent new line of research that aims to couple network theory and economics has grown in the last decade, thanks to its ability to capture information from large datasets of exports and cast it into human-readable measures to rank nations and commodities. This is not the first interdisciplinary approach to economics, as its road to complexity started many years ago. However, this new theory called *Economic Complexity* is somehow different from standard econophysics. It is a purely data-driven approach that does not aim to create a model but instead seeks to extract as much as possible information from the network of exports that at first sight may be hidden. This attracted a lot of attention at the institutional level (World Bank, UE) [6]. This research is only in its infancy, so there is still a lot of discussion, especially about what kind of information we want to extract and how.
In this thesis we will study the nations' exports from 1995 to 2019 in a bipartite network perspective, according to the economic complexity framework. We will discuss what is the essential information and how a new algorithm, based on a self-consistent use of the Shannon entropy, can enter into the theory to extract it, getting new measures of complexity of nations and products. Finally, as an original contribution, we will try to understand the dynamics of nations and forecast their future growth according to this new measure.

The Economics Road to Complexity The problem to understand why economic world trade occurs has been a central topic from the very beginning of the economic thinking. Adams Smith introduced the concept of *absolute advantage* [7]: the country that can supply the most conspicuous amount of a commodity at the cheapest price has an absolute advantage on that product. The search of this advantage was considered the force that drives countries to engage in international trade. However, a satisfying theory occurred only several decades after due to David Ricardo, who introduced the concept of *comparative advantage* [8]. With this theory we have the first example of collective behavior in economics, that emerges when nations specialize in products on which they have lower relative opportunity cost price [1], describing how countries take advantage by engaging world trade.

[1]The opportunity cost of an activity is the loss of value or benefit that would be incurred by engaging in that activity [9]

These considerations started to have mathematical foundations thanks to the initial work of the Swedish economist E. Heckscher, who introduced in 1919 [10] a theory that was formalized by his student B. Ohlin in 1933 [11]. The so called Heckscher-Ohlin model was extended and further generalized up to become the *standard economic model* [12], describing how countries reach an equilibrium trade according to their fixed non-tradable capital, such as labor and infrastructure. Further works on the model [13] showed that the accumulation of capital can cause changes in the export of a nation, both in quantity and in typology, therefore modifying the structure of the world market breaking the equilibrium of trade proposed by the Heckscher-Ohlin model.
The paradigm shift towards a theory where exports and nation's economic capitals influence each other requires a deeper understanding of their complex interaction, paving the way for the network science into economics.

We have to wait until 2003 for a first attempt to organize the world market into a network model, the so called *World Trade Web*[14], on which nations are connected through their monetary trade, setting up a model to describe the spread of economic crisis expansion among countries. This study demonstrated how this network is far from being random, starting in this sense the road to complexity into these topics.
In the same spirit, R. Hausmann and collaborators developed a model for the network of products, the *Product Space* [15], on which a proximity measure has been developed to catch the "similarity" or "proximity" among products. Countries explore this network through their production system, changing it following paths that tend to connect similar products. However, this measure of proximity, based on a conditional probability constructed on an empirical formulation of comparative advantage, does not consider asymmetrical proximity relation that can intuitively occur: for instance, oil and fruits have a clear unidirectional relation as fruits depend strongly on oil for transportation costs. These considerations evolved into a new model developed in Padua in 2016 [16], and further advanced in [17][18], on which a *gravitational model* (one of the first economic models developed by a physicist, J. Tinbergen [19] winner of a Nobel prize in economics in 1969) of trade has been used to construct a different measure of distance among products. This new network is also equipped with the notion of time, on which the dynamics of the nodes are described by a set of stochastic differential equations (SDE) that describe the evolution of exports. This model is a particular example of the interdisciplinary character of complex science, as the set of SDE used, inspired by a work of J.P. Bouchaud and M. Mézard [20] on the distribution of wealth in a society, can be transformed into a more famous equation already used to study the growth on surfaces [21], which is the *Kardar-Parisi-Zhang* equation.

Economic Complexity foundations This path of economics into complexity has recently led, thanks to the nowadays availability of large datasets of exports, to an alternative and complementary line of research called *Economic Complexity*. The aim is no longer to create a model capable to replicate the complex structure of the world trade network, but to infer as much as possible meaningful information about the nodes of the network and, possibly, use this information to deduce the future topology of the economic network. This approach lies in the more general theory of data dimensionality reduction, on which we can gain more information using a complexity approach than a mere aggregation of data. It is of course necessary to define what kind of information we want to extract from the world trade network.
Countries interact with each other by exchanging products. As C. Hidalgo wrote in [22], products are nothing else than a solidification, or crystallization, of knowledge that takes three forms: embodied knowledge in tools and material, codified knowledge in books, algorithms and formulas, and tacit knowledge also called *know-how*. While the first two types of knowledge are easily tradable, the know-how is more complex and intangible, as it is a result of a long process of

repetition, imitation and feedback. These three types of knowledge are strongly complementary: to create a house we need several materials and infrastructures (embodied knowledge), we need to know the necessary laws (codes information) and we need to know how to use the needed instruments. The more complex is a product the more knowledge is required to construct it and actually export it. In this sense production systems of nations are bounded by the amount of knowledge they embed, and is the absence of know-how, which is not easily tradable, to create this bound.
The know-how seems to be the important information we want to measure, as it is an indicator of how much a nation is potentially able to produce more complex products. As firstly noted by Adam Smith in his most famous work, the *Wealth of Nations*:

> *It is the great multiplication of the productions of all the different arts, in consequence of the division of labor, which occasions, in a well-governed society, that universal opulence which extends itself to the lowest ranks of the people*

The diversification in productions drives the growth of know-how. Specialization of individuals, that accumulate know-how, and diversification are two aspects of the same phenomenon seen from two different scale [6]. A first qualitative results of these considerations is an explanation of why for poor countries is difficult to enhance their wealth conditions, as noted by Hausmann and Hidalgo [23] and paraphrased in [6]: "as the number of possibilities grows exponentially with the variety of elements to combine, countries with few (many) of those elements will have few (strong) incentives to accumulate more elements as they may produce few (many) new combinations".
Diversification in productive systems seems to be the right proxy to infer the amount of know-how and unveil the *complexity* of nations. It is clear that a mere count of products or a naive indicator of the distribution of exports of a nation is not enough to achieve this result. The intrinsic complexity of the world trade has to be considered as well as the heterogeneity of countries and the complexity of products. A complementary measure of the complexity of nations regarding products is necessary to the task.

The first attempt to establish a measure of complexity of nations is due to C. Hidalgo and R. Hausmann in 2009 [24]. Laying on the Ricardian's idea of comparative advantage, they considered a country a producer of a product if it had a *revealed comparative advantage* (RCA) on it. The more common, but not unique [25], way to define mathematically the comparative advantage is to use the *Balassa index* [26], which is the most common name of RCA: in this approach a country is considered to have a comparative advantage on a product if its fraction in the export basket is greater than the same fraction of an "average" country, setting in this sense a threshold criterion[2]. Using RCA they constructed a *bipartite network* in which countries are connected to the products on which they revealed a comparative advantage. The graph was therefore equipped with a bi-adjacency matrix on which the elements could assume only binary numbers (0 if the country was not a relevant exporter of a product, 1 otherwise).
To define weights to each nodes according to their importance in the graph, they developed an iterative algorithm called *Method of Reflections*, obtaining from it the *Economic Complexity Index* (ECI) and the *Product Complexity index* (PCI). This was the start of the economic complexity framework and further works, using this approach, showed how to establish correlation with green economy [27], income inequality [28] and health indicators [29] for some examples.
However, this algorithm has been criticized especially for the interpretation of generalized diversification that the authors give to ECI. It was realized in [30] that this measure is orthogonal

[2] A country c has a revealed comparative advantage on a product p if and only if $RCA_{cp} > 1$. However, one could think to change the threshold from 1 to T getting $RCA_{cp} > T$. The algorithm in question showed stability on this choice.

to a bare measure of diversity (the simple count of products exported with comparative advantage) rising legitimate suspicious about the real meaning of the measure. This problem of interpretation is the first motivation on why we will discuss a new method in this thesis.

In 2012 L. Pietronero and his group developed a new method to estimate complexities among the nodes of a bipartite network, the *Fitness Method* [31], getting a new measure with apparently no problem of interpretation. Remarkably, they found that reordering the bi-adjacency matrix according to the fitness rank shows a nested structure[3], very similar to the ones observed in ecological mutualistic networks [1]. This feature in the bi-adjacency matrix indicates a hierarchical structure based on the number of products a country exports with a comparative advantage, corroborating the underlying idea that nations tend to diversify instead of specialize in few products. But if on the one hand we have a meaningful measure, on the other the non-linear nature of the Fitness algorithm gives problems of convergence, in particular when we have to deal with niche products or we change the criterion to establish comparative advantage [32]. Despite these problematics, the algorithm works if we consider aggregated products and we cut the iteration at a certain point, so a lot of ink has been spilled in this direction [33][34]. Studying the dynamics of this complexity measure unveiled heterogeneous patterns of countries' evolution [35], individuating regions of high and low predictability of growth according to fitness. From this consideration they developed an algorithm (the *Selective Predictability Scheme bootstrap*) [36] to forecast GDP, based on the *Method of Analogues* developed my E.N. Lorenz [37] in the context of atmospheric prediction. The problem of convergence in the iterative algorithm for measuring fitness is the second motivation that led to the development of the new measure discussed in this thesis. We will see that this new measure will also lead to some of the main results discussed above.

What is in common between ECI and Fitness is that the information on exports is always initially pre-processed by only considering revealed comparative advantage and, a part an exception [31], the bipartite graphs only contemplate information about whether a country is a competitor or not, regardless the amount of products it exports. Besides the huge loss of information, if the main task is to sort of "count" the number of products a country produce, it seems rather unclear why we have to only consider products on which it is revealed a comparative advantage. Moreover, as found in [25], Balassa index has shown some shortcomings to truly representing the comparative advantage, opening a discussion on other empirical measures besides RCA. More explanatory is the example of the clothing export [25]. In 1996 Italy showed an RCA of 11.4 while Germany only 0.06, with RCA criterion we would end up with the fact the Germany does not export clothes, hence it does not have the know-how related to that product: this seems rather unjustified. Moreover, RCA is a source of noise in the dynamics of the bipartite graph as, especially for least developed countries, RCA values can oscillate around the threshold [36](Supplementary Information). In the same paper, regarding the Fitness algorithm, it was proposed an alternative criterion that gets rid of thresholds, which was a problem for that algorithm [32]. They decided to model the RCA time series as the emissions of the Hidden States of an Hidden Markov Model. A first attempt to not use RCA threshold criterion is found in [38], on which an analysis with principal component reduction and machine learning techniques is proposed on large aggregation in products.
RCA is not the only cause of information loss. Export datasets provide products' classification at the finest possible level. However, most of the analysis of economic complexity concern macro-categories of products obtained from a mere aggregation of exports, in contrast with the dimensionality reduction thinking, as this approach does not preserve all the information

[3] A bipartite graph has a nested structure if its bi-adjacency matrix is comparable to an upper triangular matrix, after a proper reorder of rows and columns.

embedded at the finest level. RCA and product aggregation cause a loss of information that has never been addressed in the literature.

A novel approach that makes full use of the information contained in the export datasets started to grow in [18], but we had to wait 2021 for a full description of the method [39] (which is the starting point of the thesis). Within this approach, the focus has moved from an iterative counting of competitive products (ECI and Fitness) to a non-trivial measure of diversification in the export's basket. The measure is based on a self-consistent use of the Shannon entropy, which is a common and universally accepted indicator of diversity, and shows an exponentially fast local and global convergence to a unique fixed point. This is the first attempt in the economic complexity literature to perform dimensionality reduction without considering the comparative advantage of products, making this measure free on any data-preprocess assumptions. Moreover, Shannon entropy allows performing coarse-grained analysis of nations and products going beyond simple data aggregation of exports, opening a complete novel analysis in the economic complexity framework according to the dimensionality reduction thinking. The proven convergence of the algorithm, the clear interpretation of the measures, its complete use of the information embedded in the datasets, and its not dependence on conceptual data-preprocessing, make this measure a good candidate to rank nations and products according to their relevance in the trade market. In addition, the stability of the algorithm and its use in weighted bipartite graphs inaugurate a unique interdisciplinary approach to network science going beyond economics.

Structure of the Thesis This thesis will start with presenting the dataset used and characterizing different ways to define a complex bipartite network to model the worldwide export during 2019. After a brief introduction on revealed comparative advantage, ECI and Fitness algorithms, we will highlight the importance of diversification as the essential ingredient in the economic complexity framework and how entropy emerges as a natural candidate to measure it. A technical introduction to the self-consistent iterative scheme developed in [39] will be the next step, accompanied by a depth study of local and global convergence of the algorithm. We will observe how this algorithm returns nations' complexity measures that establish ranks in line with the economic narrative and products' ubiquity measures that discriminate according to their importance in the world market. These results will be enhanced by a coarse-grained analysis in product categories, showing a novel economic complexity approach that was impossible with ECI and Fitness. To conclude the first chapter, we will discuss the role of revealed comparative advantage and how its adoption would affect the entropy complexity measure.
The dynamics of the nations' complexity measures will be presented in the second chapter, in the same fashion of [35]. We will couple the entropy to a monetary measure, constructing a bidimensional economic plane tracking the counties' time series from 1995 to 2019. A coarse grain technique will reveal a flow structure on this plane, showing different dynamical patterns in the macroeconomic landscape. As the main result, we will demonstrate how entropy discriminates among countries according to the stability of their economy and their possibility of growth. We will also analyze the role of revealed comparative advantage in this dynamical context, observing that it captures different dynamics.
Finally, in the last chapter, we will use the dynamics to forecast GDP growth using the Selective Predictability Scheme [36]. The idea is to look at historical dynamics of nations with comparable entropy and GDP (Gross Domestic Product) to infer future growth. However, in the original formulation of the algorithm, the problem to choose the right "comparable" nations' dynamics was not addressed. Therefore, we will individuate and solve this problem using a statistical learning approach to historical data, combined with a kernel regression. The algorithm's accuracy will be compared to the International Monetary Fund's (IMF) predictions, and the improvement will corroborate our choice of entropy as a good candidate to measure nations' complexity.

List of Abbreviations and Symbols

Abbreviations

HS	Harmonized System	5
RCA	Revealed Comparative Advantage	6
GDP	Gross Domestic Product	16
GDPpc	Gross Domestic Product per capita	16
GDPpcPPP	Gross Domestic Product per capita in Purchasing Power Parity	34
IMF	International Monetary Fund	45
SBSb	Selective Predictability Scheme bootstrap	45
cSBS	Convergent Selective Predictability Scheme	50
CAGR	Compound Annual Growth Rate	51
MAE	Mean Absolute Error	51
RMSE	Root Mean Square Error	58

Most Used Symbols

X_{cp}	Bi-adjacency matrix	6
M_{cp}	Bi-adjacency matrix binarized with RCA	7
ξ_{cp}	Weighted export basket of a country c	16
ζ_{cp}	Weighted country's share of a product p	16
H_c	Entropy of a nation	16
H_p	Ubiquity of a product	16
r_S	Spearman correlation coefficient	16
$\mathbf{x}_{c,t}$	Point in the entropy-income plane, related to a country c at a time t	31
$\mathbf{v}_{c,t}$	Velocity or trend related to a country c at a time t	32
σ_b^2	Measure of chaos in a coarse-grained box	33
$\mathbf{x}_{\tilde{c},\tau}$	Analogue of $\mathbf{x}_{c,t}$ in the entropy-GDP plane	46
$\delta\mathbf{x}_{c,t}$	5 year displacement of $\mathbf{x}_{c,t}$	46
$\mathbf{H}$	Covariance matrix or bandwidth of a multivariate Gaussian p.d.f.	47
$r(\mathbf{x}_{c,t}\,,\,\mathbf{x}_{\tilde{c},\tau})$	Mahalanobis distance of two points in the entropy-GDP plane	48
E	Error in forecasting GDP growth using CAGR	51

Contents

Economic Complexity

Economic complexity is a new line of research that aims to measure the underlying intangible capabilities of nations with complexity measures computed by analyzing the exports properly. In the literature, two main algorithms can be found, economic complexity index [24] and fitness [31], but the first does not have a clear interpretation [30], while the second shows problems of convergence [32]. In this chapter, a new complexity measure will be introduced, based on the work [39], and a static analysis of the year 2019 will be presented to support the work already done. An original discussion about the role of revealed comparative advantage and different editions of the harmonized system will be presented.

1.1 Dataset and Aggregation

The Dataset we are using is the BACI supplied by CEPII (it gathers and harmonizes different declarations in monetary exports listed by COMTRADE) that covers exports and imports in the world trade network from 1995 to 2019. The products are classified using different Harmonized System (HS) editions, but for this particular analysis about 2019, we decided to use the most recent, the 2017 edition (HS17)[40]. The classification works as a 6 digits code: the first two digits designate the HS chapter, the second two digits label the HS heading, and the last two indicate the HS subheading. We propose an example of how the HS database is organized:

- 01 - **Live animals**
  - 0101 - **Live horses, asses, mules and hinnies**
    - 010121- Pure-bred breeding animals
    - 010129 - Other
    - 010130 - Asses
  - 0102 - **Live bovine animals**
    - 010221- Pure-bred breeding animals
    - 010229 - Other
  - ...
- ...
- 72 - **Iron and Steel**
  - 7201 - **Pig iron and spiegeleisen in pigs, block or other primary forms**
    - 720110 - Non-alloy pig iron containing by weight 0.5% or less of phosphorus

        * 720120 - Non-alloy pig iron containing by weight more than 0.5% of phosphorus
        * 720150 - Alloy pig iron; spiegeleisen
    - ...
- ...

This classification offers a naturally aggregation of products in chapters, headings, and sub-headings, for about 5400 products in total. Indeed, this kind of classification was helpful in many works found in the economic complexity literature, which usually use products' classification at 4 digits level. We can aggregate the BACI dataset in a more simple form, creating the bi-adjacency matrix $X_{cp}(t)$ where the elements represent the total export of a product p done by a country c in a given year t. In the whole chapter, we will omit the time dependence of the bi-adjacency matrix as we will only deal with the exports made during 2019. In this scenario, this matrix allows us to represent the world trade network of exports as a complex undirected bipartite network, where in the first layer we have countries while in the second one we find products. The links are the elements of the matrix X_{cp}.

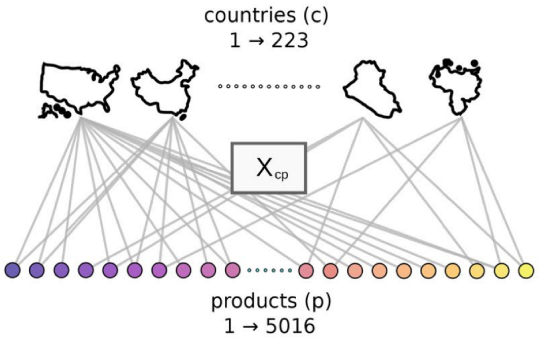


Figure. 1.1: Figure took from [39]

It is common practice to use binarization methods in bipartite networks to aggregate data that exceed a certain threshold. In this way, one can think to gather information only from the most important links of the network. For the economic complexity approach, the most common criterion is based on *Revealed Comparative Advantage* (RCA) [15, 31] that indicates if a country can be considered an effective producer of a specific product. This idea is based on Ricardian's works of comparative advantage; RCA, also called *Balassa index*, is nothing else than an empirical construction to measure it.

$$RCA_{cp} = \frac{X_{cp} / \sum_{p'} X_{cp'}}{\sum_{c'} X_{c'p} / \sum_{c'p'} X_{c'p'}} \tag{1.1}$$

In the RCA criteria, a country has to have an $RCA_{cp} \geq 1$ to be accepted as an important producer of a product p. Intuitively, with a threshold equal to one, we are saying that a country is an effective producer of a product p if its export share of that product $X_{cp} / \sum_{p'} X_{cp}$ is bigger than the average export share made by all countries $\sum_{c'} X_{c'p} / \sum_{c'p'} X_{c'p'}$.
In this way, one can construct a binarized version M_{cp} of the matrix X_{cp} in which we have only the information whether a country is a competitor of a specific product or not. A filtered matrix $\tilde{X}_{cp}$ can be constructed as well, where remains the information of the amount of product that

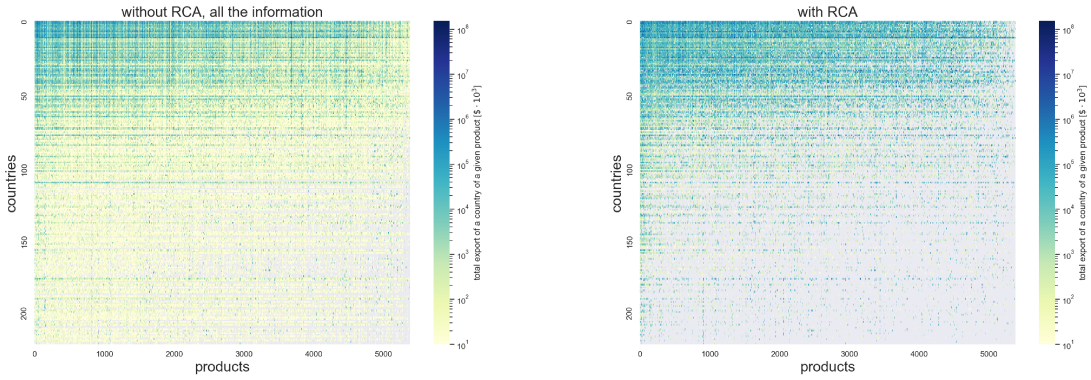


Figure. 1.2: In the left figure is represented the biadjacency matrix X_{cp} sorted with an increasing country entropy and am ascending product ubiquity, using the method that will be explained in the next section. In the right figure is represented the $\tilde{X}_{cp}$ sorted, it is visible a more evident nested structure. The data are considered for 2019

is exported.

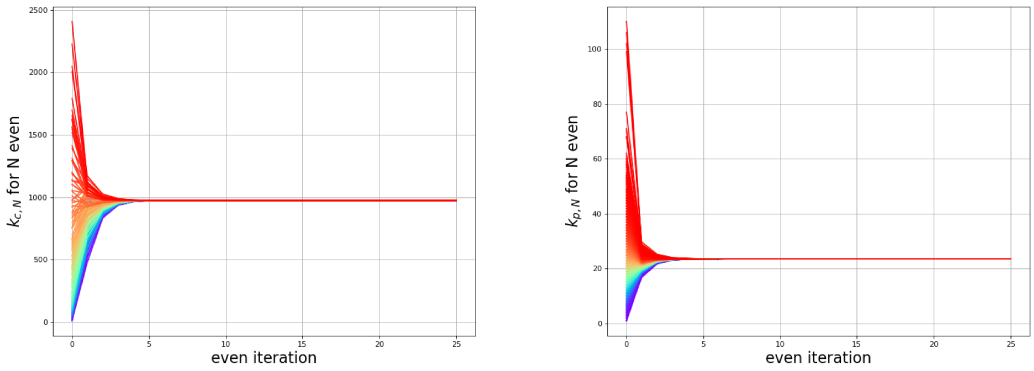
$$M_{cp} = \begin{cases} 0 & \text{if } RCA_{cp} < 1 \\ 1 & \text{if } RCA_{cp} \geq 1 \end{cases} \qquad \tilde{X}_{cp} = \begin{cases} 0 & \text{if } RCA_{cp} < 1 \\ X_{cp} & \text{if } RCA_{cp} \geq 1 \end{cases} \tag{1.2}$$

An essential element of consistency of our method of complexity measure is that a nested structure in the bipartite network is revealed [41]. With a particular reorder, based on the ranking criterion exposed in the following sections, a nested structure emerges from the rows and columns of the bi-adjacency matrix (figure 1.2). We can observe that the RCA criterion eliminates most of the links in the yellow region of the color bar (low export), erasing almost an 18% of them. We will start with a short introduction of the two most popular algorithms to measure node's complexity. It is to highlight the fact that both use RCA, although the loss of information due to it has never been discussed in the economic complexity literature.

1.2 Ranking with ECI and Fitness

Economic Complexity Index The first attempt to create a rank using only exports information in the world trade network was made by C.Hidalgo and R. Hausmann in [24]. In this work, they treated the exports as a complex bipartite network using as links the binarized matrix M_{cp} obtained through the RCA criteria. They developed the *Method of Reflections*, called in this way because of the symmetry of the algorithm, where it is computed the average value of the previous-level properties of a node's neighbors in a linear coupled iteratively way.

$$\begin{cases} k_{c,N} = \dfrac{1}{k_{c,0}} \sum_p^{N_p} M_{cp} k_{p,N-1} \\ \\ k_{p,N} = \dfrac{1}{k_{p,0}} \sum_c^{N_c} M_{cp} k_{c,N-1} \end{cases} \tag{1.3}$$



(a) country ranks (b) product ranks

Figure. 1.3: Complexity of countries (a) and ubiquity of products (b) using th full digit HS system classification, computed using the Method of Reflection

Where N_c and N_p are respectively the number of countries and products. Using as initial condition the number of links connected to a node

$$k_{c,0} = \sum_{p}^{N_p} M_{cp} \qquad k_{p,0} = \sum_{c}^{N_c} M_{cp} \tag{1.4}$$

The first initial condition $k_{c,0}$ indicates the number of products on which the country c shows comparative advantage, while the second initial condition $k_{p,0}$ represents the number of countries having a comparative advantage on the product p. They are respectively bare measures of a country's diversification and a product's ubiquity, as they do not take into account the difference of each country and product. With the method of reflections the difference among the nodes of the network is extracted and for countries even variables ($k_{c,2}$, $k_{c,4}$, ...) are considered generalized measure of diversification, while for products even variables ($k_{p,0}$, $k_{p,2}$, $k_{p,4}$, ...) are generalized measures of ubiquity.
This simple method has an uninformative fixed point, as the rank between countries is captured in a progressively shrinking difference (fig 1.3). The reason is that $k_{c,N}$ tends to converge to a vector with all components that are the same. Although they used the 4-digit classification in the HS for products in their work, we decided to show the results for the finer classification of products, as this does not change the shape of the graphs.

The convergence is due to the linearity of the iterative map

$$k_{c,N} = \frac{1}{k_{c,0}} \sum_{p}^{N_p} M_{cp} k_{p,N-1} = \frac{1}{k_{c,0}} \sum_{p,c'} \frac{M_{cp} M_{c'p} k_{c',N-1}}{k_{p',0}} \tag{1.5}$$

for large N this iteration map is equivalent to an eigenvalue problem

$$k_c = \sum_{c'} M'_{cc'} k_{c'} \qquad \text{with} \qquad M'_{cc'} = \sum_{p} \frac{M_{cp} M_{c'p}}{k_{p',0} k_{c,0}} = \frac{\sum_p M_{cp} M_{c'p}}{\sum_{c',p'} M_{c,p'} M_{c',p'}} \tag{1.6}$$

the solution of the system are the eigenvectors of $M'_{c,c'}$, but since it is a transitional probability matrix ($\sum_{c'} M_{cc'} = 1$) its first eigenvector has equal elements.

The authors noted this problem, and in a subsequent work [42] they tried to fix it by taking the eigenvector associated to the second largest eigenvalue of the iterative map $M'_{cc'}$. They called it economic complexity index (ECI) and gave to it the meaning of a generalized measure of diversification, hence complexity. The principal issue pointed out in [30] is that the ECI vector is orthogonal to the bare diversification $k_{c,0}$, undermining its interpretation.

Fitness Method A novel approach, called *Fitness Method*, that uses a non-linear iterative algorithm, and therefore does not suffer from the eigenvector problem illustrated previously, was developed by the group of L.Pietronero [31]. In the spirit of the work done with the method of reflections, they gathered data in the binarized matrix M_{cp}, but also in the weighted matrix $\tilde{X}_{cp}$ normalized with the worldwide export of a product, using the RCA criteria.
The idea is that the fitness F_c of a country is proportional to the sum of the products with a revealed competitive export (just like the ECI), weighted by their complexity Q_p. Computing the complexity of a product is more subtle; it is inversely proportional to the number of countries which export it (the more a product has revealed competitor, the less it is complex). In addition, if a country has a high fitness this should reduce the weight in bounding the complexity of a product, and the countries with low fitness should strongly contribute to the bound on Q_p. This is a consequence of the nestedness of the bipartite graph, countries with high fitness tend to export everything. This idea is written mathematically in the following algorithm

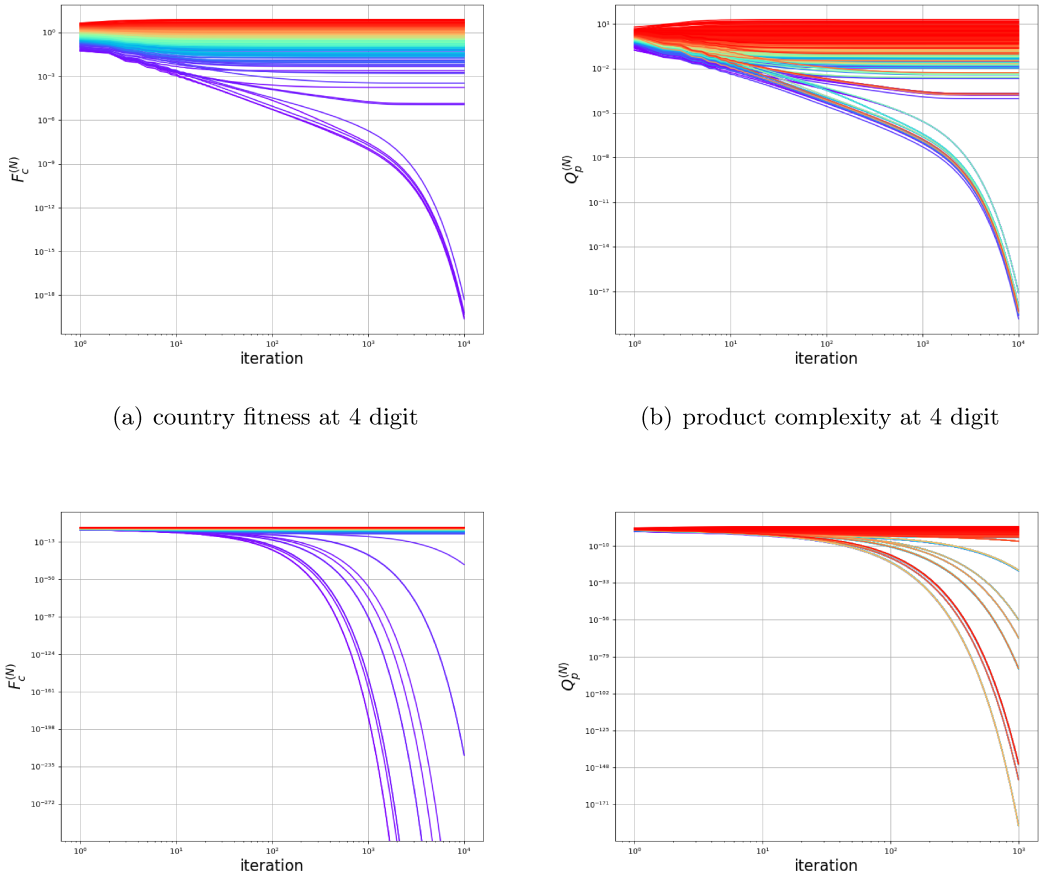
$$\begin{cases} \tilde{F}_c^{(n)} = \sum_p^{N_p} M_{cp} Q_p^{(n-1)} \\ \\ \tilde{Q}_p^{(n)} = \dfrac{1}{\sum_c^{N_c} M_{cp} \frac{1}{F_c^{(n-1)}}} \end{cases} \quad \rightarrow \quad \begin{cases} F_c^{(n)} = \dfrac{\tilde{F}_c^{(n)}}{\left\langle \tilde{F}_c^{(n)} \right\rangle_c} \\ \\ Q_p^{(n)} = \dfrac{\tilde{Q}_p^{(n)}}{\left\langle \tilde{Q}_p^{(n)} \right\rangle_p} \end{cases} \tag{1.7}$$

The initial condition are $F_c^{(n)} = 1$ and $Q_p^{(n)} = 1$ for all countries and products. The analysis in [31] was made using the 4 digits classification of the HS for products, and apart from a few nodes that converge to zero, the algorithm stabilized at a fixed point. As noted in [32], the problem arises when the algorithm has to deal with niche products. If we use the 6 digits classification in the HS for products, the algorithm shows convergence problems (see figure 1.4). It is still possible to get a rank and a measure by stopping the algorithm at a certain point, but this is not a good solution since it adds an arbitrariness leading to different fitness and complexity values.

Diversity is the essential ingredient in both approaches, as the more a country has revealed comparative advantage in different products, the greater its complexity. In the next section, we will deeply analyze the role of diversification and see how entropy emerges as a natural candidate for this task.

1.3 The Role of Diversification

Diversification at the country level seems counter-intuitive, as one could think that countries will reach different levels of specialization in a free trade market according to their know-how. R. Hausmann wrote about this in an opinion article *"Many believe that cities, regions, and countries should specialize: they cannot be good at everything, so they must concentrate on their comparative advantage. But, while this idea seems obvious, it is both wrong and dangerous [43]."* It is misleading to equate the benefits of individual specialization with those of specialization at a larger scale. Specialization at the individual level is needed to



(a) country fitness at 4 digit

(b) product complexity at 4 digit

(c) country fitness at 6 digit

(d) product complexity at 6 digit

Figure. 1.4: Fitness of countries and complexity of products computed using the Fitness Method. There are problems of convergence in (c) and (d) when the algorithm has to deal with niche product (products that are exported from very few countries in very low quantity). Using the 4 digit classification we remove a little niche products from the bipartite network, however there are still some problem of convergence of the iterative map.

gather the necessary know-how to create something, but at the scale of cities or countries, this individual specialization yields diversification. If the opposite were true, we would observe in the bi-adjacency matrix X_{cp} a block diagonal shape, but we have seen in figure 1.2 that the bi-adjacency matrix shows a clear nested structure, after a proper reorder. In other words, wealthy countries tend to export all the possible products. In this sense, diversification is the most critical indicator for the economic complexity approach: it drives economic development, as also indicated in other works [38].

Both Fitness and, according to the authors, ECI are complex measures of diversification constructed on a non-trivial count of products on which nations revealed comparative advantage. However, this concept of diversification does not contemplate the relative weights that each production has on the productive system. A more intuitive way to observe the diversification of nations is to look at how their products are distributed in their export baskets, hence at the quantity $\xi^0_{cp} = X_{cp}/\sum_p X_{cp}$. The most diversified country should have a narrow distribution peaking close to the equal share value $1/N_p$ (where N_p is the number of product trade in the world trade network that year), getting broader with countries with low diversification. Using

the BACI dataset, we can get some insights into this role of diversification. According to World Bank classification [44] we take four different countries (Figure 1.5): Sudan (SDN) as a low income economy, Iran (IRN) as a lower-middle income economy, Brazil (BRA) as a upper-middle income economy, and finally Italy (ITA) as an high income economy. Sudan shows a broader peak far from the equipartition value; for Iran and Brazil, the peak moves towards equipartition and their distribution gets narrower. Italy shows a more narrow distribution with a peak rather close to the equipartition value.

Diversification in the export shares already embeds some information about the wealth of a nation. It is fascinating that the median of the basket shares of a country, which can be seen as a naive measure of diversification, correlates very well with the total export of countries that export less than 1000 products, see figure 1.6 (a). This line individuates a set of underdeveloped countries (such as Iraq, Venezuela and Sudan), but also not underdeveloped ones like San Marino (SMR) or Andorra (ADR), that have only the constraint of being a very small exporters due to their dimension. Future diversification measures should capture this *poverty line* with also the ability to exclude small countries with a high total export per capita. However, the poverty line is no longer visible when we plot against export per capita (see figure 1.6 (b)). This problem is resolved using a more complex diversity indicator, as we will see in the results section.

In the next section, we will introduce a new method developed by the group of Padua [39] that aims to measure this diversification correctly, using the machinery of information theory.

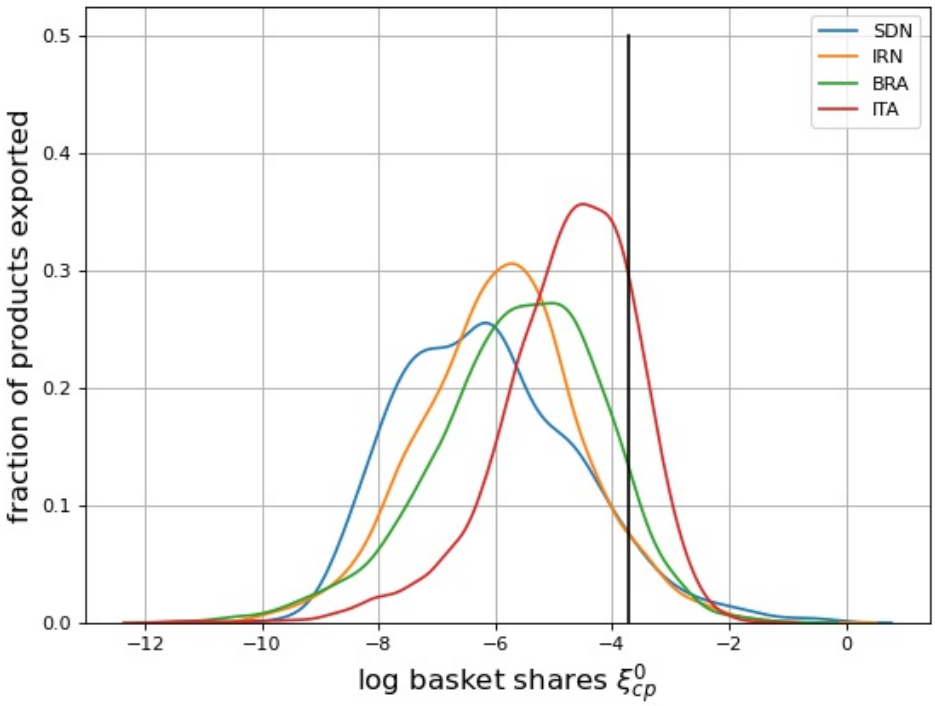
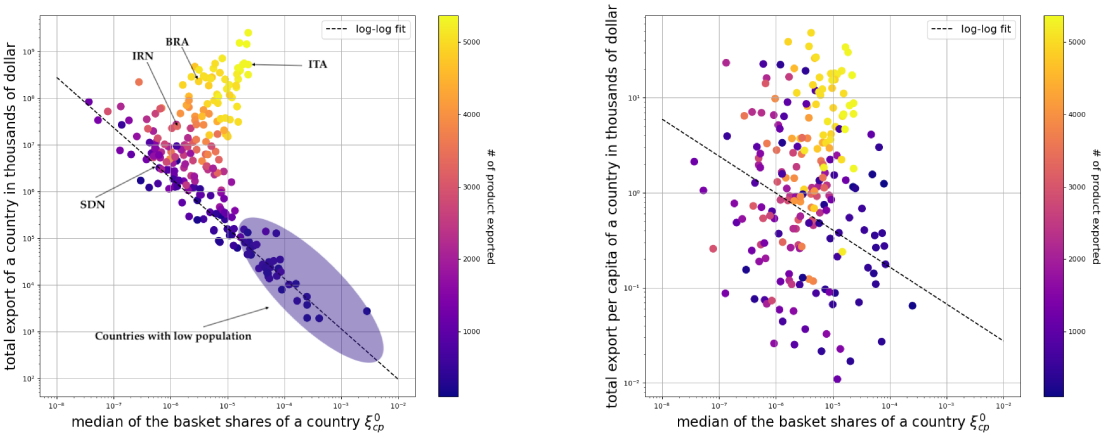


Figure. 1.5: In these plots, we observe Gaussian kernel density estimation with a bandwidth equal to one of the diversification of the export shares ξ^0_{cp}. Four different countries are considered, according to their income classification. Differences in these densities are present among these main sets. The vertical line represents the ideal equipartition value.



(a) (b)

Figure. 1.6: Figure (a) depicts a scatter plot of export - median of basket share. A clear correlation is visible with countries which export less than 1000 products. In figure (b), we changed the y-axis with export per capita. As a result, a region or line with countries that export a few products is no longer visible. As for some countries, population information is lacking, few nations are depicted in the (b) plots.

1.4 Entropic Complexity Measure

A natural candidate function to measure diversification is the Shannon Entropy: given a set of probabilities $\{p_i\}_{i=1,\dots,N}$ and $\sum_i p_i = 1$ the Shannon entropy is defined as

$$H = -\sum_{i=0}^{N} p_i \log(p_i) \tag{1.8}$$

We can think of the probabilities as the relative occupation of a collection of N states, hence the shares of a country's export basket. As pointed out before, we can define the share ξ^0_{cp} of the product p in the basket of country c as the fraction of the product's export in the overall export of that country.

$$\xi^0_{cp} = \frac{X_{cp}}{\sum_{p'} X_{cp'}} \tag{1.9}$$

In the same way, we can define the export share ζ^0_{cp}, which is the fraction of the export p of the country c in the overall export of that same product at a global level.

$$\zeta^0_{cp} = \frac{X_{cp}}{\sum_{c'} X_{c'p}} \tag{1.10}$$

The apex zero indicates that those quantities are the basis of our construction of the measure. Indeed, these shares only gather monetary information and have no insight into the relationship between different products or nations. The bare entropic and ubiquity indicator can be defined using these shares as probabilities, as they are already normalized and sum to one.

$$H^0_c = -\sum_{p}^{N_p} \xi^0_{cp} \log(\xi^0_{cp}) \qquad H^0_p = -\sum_{c}^{N_c} \zeta^0_{cp} \log(\zeta^0_{cp}) \tag{1.11}$$

Where N_p is the total number of products traded, and N_c is the total number of countries considered. This entropy has not to be intended as an indicator of the diversity of the p.d.f. in

figure 1.5, indeed one can argue that an ideal diversified country should have a delta distribution centered on the equipartition line. The entropy is instead an indicator of the diversity of the shares ξ^0_{cp} as a function of products.
As mentioned before, these indicators do not have the information of different products and nations: for example, we could change crude oils with copper in the basket of a country's export ξ^0_{cp} and this would not change the bare entropic indicator. In the same way, we could change USA with Tunisia for computing the export share ζ^0_{cp} of a product, and this would not change the result.
The distinction among the nodes of the bipartite network is obtained with a self-consistently re-weighting of the share matrices, in the same fashion as Fitness and ECI algorithms. Using a coupling approach: the more a product is ubiquitous, the less it should contribute to the diversification of a country (i.e. to the entropy of a country), while the more a country is diversified (in our approach means wealthy), the less it should contribute to determining the ubiquity of a product. We adopt the idea, corroborated by the proven nested property of the graph, that wealthy and developed countries have a consolidated diversification and export all kinds of goods. In this sense, diversified (wealthy) countries should contribute less than poor diversified nations to the ubiquity of products.

For instance, if USA exports a lot of a product called P that is not exported by a large share of the less diversified competitors, the product P will end up with a small ubiquity indicator. On the other hand, if the same product P is exported by many countries with a small entropic indicator, hence with low diversification, the product will end up with a high ubiquity indicator. USA has a high entropic indicator, so it can produce and export a vast quantity of diversified products whether they are complex or not, while countries with a small entropic indicator produce and export only products that are not complex. We can think of the ubiquity measure as the opposite of the complexity of a product [39], in a sense given by the fitness algorithm.

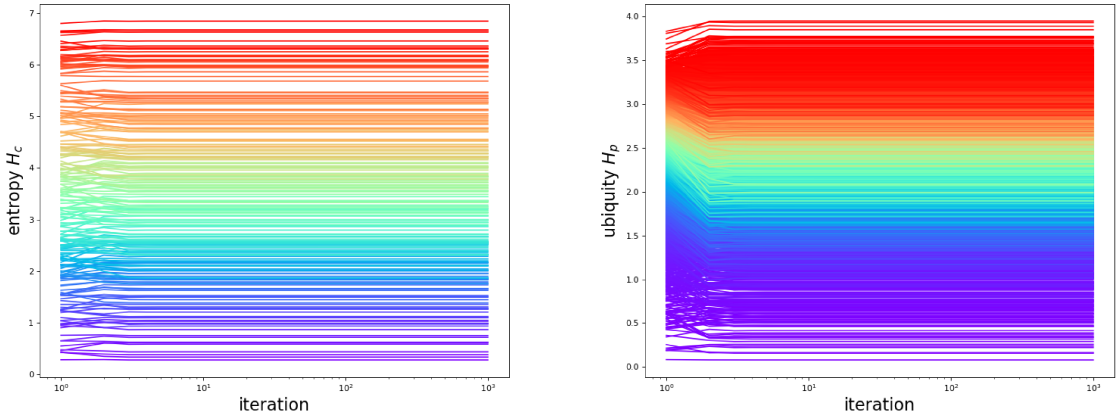
This idea was proposed mathematically using an iterative algorithm with initial conditions the bare entropic indicator (1.11) [39]. In contrast with Fitness and ECI, we iteratively re-weight the shares instead of the measures with this approach.

$$\begin{cases} H_c^{(k+1)} = -\sum_p^{N_p} \xi_{cp}^{(k)} \log(\xi_{cp}^{(k)}) \\ \\ H_p^{(k+1)} = -\sum_c^{N_c} \zeta_{cp}^{(k)} \log(\zeta_{cp}^{(k)}) \end{cases} \tag{1.12}$$

with shares at the k-th step defined as

$$\xi_{cp}^{(k)} = \frac{X_{cp} f\big(H_p^{(k)}\big)}{\sum_{p'} X_{cp'} f\big(H_{p'}^{(k)}\big)} \qquad\qquad \zeta_{cp}^{(k)} = \frac{X_{cp} g\big(H_c^{(k)}\big)}{\sum_{c'} X_{c'p} g\big(H_{c'}^{(k)}\big)} \tag{1.13}$$

The functions f and g are the weighting functions that re-weight the share matrices at each iteration. We want to give higher contributions to products with low ubiquity in computing the entropy of a country, and in the same way, we want to give more weight to countries with low entropy to compute the ubiquity of a product. We need two functions f and g that invert the concept of entropy and ubiquity, respectively.
There are many ways to achieve this, but a simple way is to take the complementary of the entropy. The Shannon entropy stays in a compact set because it is a continuous function bounded from below by 0 that indicates maximum information on a stochastic variable, hence a minimum diversification (a delta probability distribution, hence a monopoly of one product), and also bounded from above by $\log(N)$ indicating minimum information, therefore a maximum



(a) (b)

Figure. 1.7: Entropy (ubiquity) of each country (product) after 10, 100, 1000 iterations. We can clearly see that after a very small transient the iteration process stabilized. In figure (b) is more evident a reorganization in the ranking of the products than the entropy case due to the bigger initial vector H_p^0.

diversification (a uniform probability distribution). Leveraging on this property, we can define the re-weighting function as

$$f(H_p) = \log(N_c) - H_p \qquad g(H_c) = \log(N_p) - H_c \tag{1.14}$$

1.4.1 Fixed point analysis

The scheme of (1.12) can be seen as a closed map φ between compact sets.

$$\varphi: \quad [0, \log(N_p)]^{N_c} \otimes [0, \log(N_c)]^{N_p} \quad \rightarrow \quad [0, \log(N_p)]^{N_c} \otimes [0, \log(N_c)]^{N_p} \tag{1.15}$$

This feature is guaranteed by the boundedness of the Shannon entropy, which also compares in the weighting functions. These properties allow us to use the Brouwer's fixed point theorem [45] that ultimately prove the existence of a fixed point $\{H_c, H_p\}_{p=1,\dots,N_p}^{c=1,\dots,N_c}$ for the map φ.
We can see from a simple line plot that this algorithm is fast convergent and is stable (see figure 1.7)

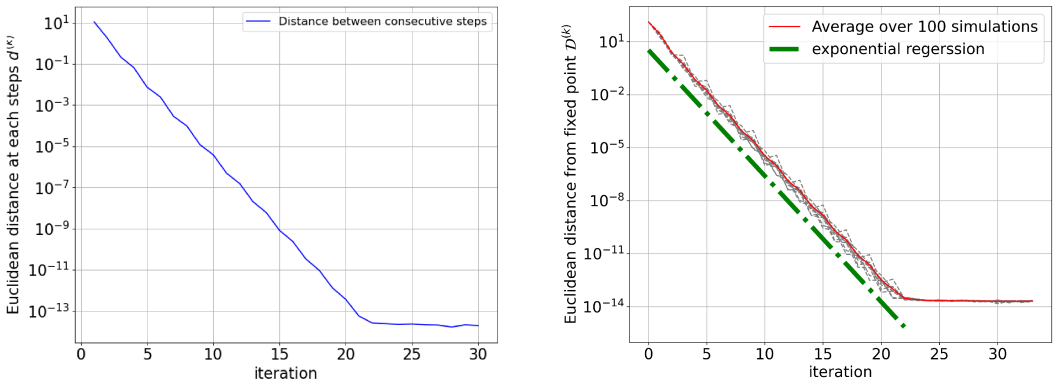
Local convergence To study the algorithm's convergence, we can compute the Euclidean distance between consecutive steps of the iteration process.

$$d^{(k)} = \left(\sum_c \left(H_c^{(k+1)} - H_c^{(k)} \right)^2 + \sum_p \left(H_p^{(k+1)} - H_p^{(k)} \right)^2 \right)^{1/2} \tag{1.16}$$

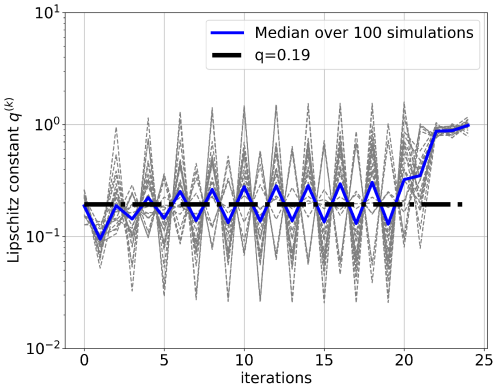
The algorithm reaches a fixed point exponentially fast, as it reaches the double-precision slightly above 20 iterations, figure 1.8(a).

Global convergence We can also check numerically if this algorithm is globally convergent by iterating the scheme for randomly initial conditions. In particular, we take random initial conditions from uniform distributions to simulate different initial bare entropies.

$$H_c^0 \sim \mathcal{U}[0, \log(N_p)] \qquad H_p^0 \sim \mathcal{U}[0, \log(N_c)] \tag{1.17}$$



(a) (b)



(c)

Figure. 1.8: Study of the convergence of the algorithm: In figure (a), to study the local convergence, we plot the Euclidean distance $d^{(k)}$ evaluated between consecutive steps; it clearly shows a fast exponential convergence before reaching the double precision at the 23rd iteration. To study the global convergence we plot, in figure (b), the Euclidean distance at each step of the simulation from the fixed point obtained after 100 steps of the iteration process, while the gray dotted line represents the first 10 simulations. We get from the graph an exponential convergence until the line flattens when it is reached double-precision after the 23rd iteration step. The exponential decay coefficient $D^{(k)} \sim e^{\alpha k}$ is computed as $\alpha = -1.54$. Figure (c) represents the evolution of the Lipschitz constant, which after 23 iteration we pass the double precision, they gray lines indicate 10 simulations.

For each simulation, we compute at each step the Euclidean distance from the fixed point, obtained from data after 100 iterative steps with the bare entropic indicator as initial conditions.

$$\mathcal{D}^{(k)} = \left(\sum_c \left(H_c^{(k)} - H_c \right)^2 + \sum_p \left(H_p^{(k)} - H_p \right)^2 \right)^{1/2} \tag{1.18}$$

The graph in figure 1.8(b) shows an exponential decay of this distance $D^{(k)} \sim e^{\alpha k}$ and a curve parametric fit allows us to compute the coefficient $\alpha = -1.54$. With this information, we can compute the Lipschitz constant q [46], an indicator of how good is the global convergence

$$q = \lim_{k\to\infty} \frac{\mathcal{D}^{(k+1)}}{\mathcal{D}^{(k)}} = \lim_{k\to\infty} \frac{e^{\alpha(k+1)}}{e^{\alpha k}} = e^{\alpha} \tag{1.19}$$

A Lipschitz constant $q = 0.19$ indicates that the algorithm is globally convergent [4]. We can observe the evolution of the Lipschitz constant in figure 1.8(c).
The iterative scheme converges to the same fixed point, for which these consistency relations hold ((1.20) and (1.21))

$$\begin{cases} H_c = -\sum_p^{N_p} \xi_{cp} \log(\xi_{cp}) \\ \\ H_p^{=} - \sum_c^{N_c} \zeta_{cp} \log(\zeta_{cp}) \end{cases} \tag{1.20}$$

Using a direct re-weight of the shares, instead of the measure, allows to define weighted shares normalized concerning country's exports, which are endowed of the different information that each node in the bipartite network has.

$$\xi_{cp} = \frac{X_{cp} f\big(H_p\big)}{\sum_{p'} X_{cp'} f\big(H_{p'}\big)} \qquad\qquad \zeta_{cp} = \frac{X_{cp} g\big(H_c\big)}{\sum_{c'} X_{c'p} g\big(H_{c'}\big)} \tag{1.21}$$

The fixed point of the iterative scheme are generalized measure of diversification of the export shares, as they embed the information of the non-trivial topology of the complex bipartite network.

1.5 Results

In this section we are going to use also other information taken from World Bank datasets, such as: population [47], GDP [48] and GDP per capita [49].

To observe how entropy discriminates among countries, we plot the fixed point of the iterative scheme, computed after 100 iterations, against some monetary indicators. Of course, the most used monetary indicator is the Gross Domestic Product (GDP), which measures the market value of all the final goods and services produced in a specific period, however it is criticized by some economist who think that it is a wrong tool for measuring well-being and sustainability [50]. In this sense, we also consider the total export of a country $\mathcal{E}_c = \sum_{p'} X_{cp'}$. This monetary measure is simpler to compute than GDP and has the advantage that it can be obtained directly from the BACI dataset. Moreover, export and GDP correlate very well (see figure 1.9), so we can use the total export as a proxy of GDP, as done in [39]. The correlation is evaluated using the Spearman correlation coefficient.

Spearman correlator coefficient This coefficient is defined as the Pearson correlation coefficient between the rank variables [51]

$$r_S = \frac{\mathrm{Cov}\big[\mathrm{R(X)},\, \mathrm{R(Y)}\big]}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{1.22}$$

Where R indicates the rank function, a map that returns a vector of ordered integer numbers. It is a nonparametric measure of rank correlation and it asses how well the relationship between two variables can be described using a monotonic function. In this context, where we are interested in correlations beyond a simple linear model, this coefficient is more appropriate than a Pearson correlation coefficient.

Another interesting correlation is between entropy and population. We observe that they do not have a strong correlation (see figure 1.9), meaning that entropy does not depend on the

[4] $q < 1$ indicates that the map associated with the algorithm is a contraction.

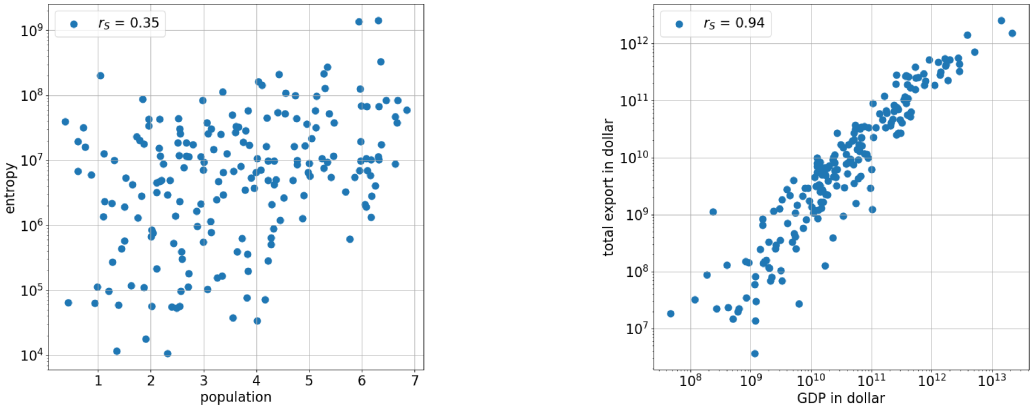


Figure. 1.9: In the left graph, we plotted entropy against the population, is visible that these two quantities do not have a strong correlation and in fact the Spearman coefficient turns out to be $r_S = 0.34$. In the right graph, we plotted the total export against GDP, these measures correlate very well with a Spearman coefficient of $r_S = 0.94$.

population of countries. However, it is natural to think that the possibility of diversifying an economy depends on how many people participate in it, and indeed, there is a small positive correlation ($r_S = 0.35$). This consideration yields us to take into account also intensive monetary measures such as GDP per capita (GDPpc) and total export capita.

Nations and entropy In figure 1.10 we observe the correlation between entropy and extensive monetary indicators (total export and GDP). There is a relatively high and positive correlation ($r_S \simeq 0.6$), but more importantly, entropy discriminates very well among countries that export a different number of goods. In figure 1.11, studying the correlation with intensive monetary measures, we observe a smaller but still positive correlation ($r_S \simeq 0.4$). According to the number of products they export, the distinction among countries is still present but less evident, suggesting classifications of nations on the entropy-GDPpc plane. This consideration will be the central point in the second chapter, where we will try to observe macroeconomic dynamical patterns by coupling entropy with intensive monetary indicators. A macro region of countries that export few products is highlighted, and remarkably, countries like San Marino (SMR) and Andorra (AND) move away from it. Other countries that have small entropy but still high GDPpc are basically of two types: small countries that are very specialized (Bermuda, Cayman Island, etc.), and one-product dependent countries, like Qatar or Saudi Arabia.
It is peculiar the position of India in the entropy-GDPpc plane, it has an high entropy but still a low GDPpc, similarly to what China was two decades ago. This type of countries are characterized by a stable economic growth, and this will be proved in the second chapter.

Products and ubiquity The only monetary information that we can have for products is the total export of the product made by all countries[5]. In figure 1.13 a scatter plot between ubiquity and worldwide export is presented. We find widespread goods such as copper, oils, and cement in the upper-right plane. In the upper-left plane, we find products with a high global export but not produced by so many countries; these products are not simple to produce, like telephones, automatic data processing machines, wristwatches, or palm oils which depend on the climatic condition of nations. In the lower-right plane, we find standard products traded by

[5]Exists other monetary measures that can be related to products, such as PRODY or Sophistication [33]

a few countries; a high ubiquity indicates that these products are exported by a vast range of low entropic countries. In the lower-left plane, we find unique products like spent fuel elements of nuclear reactors that are exported in small quantities by a few countries, but also we can find old products that are still exported by a small percentage of nations like cathode-ray television tubes. In figure 1.21, in the next section, we will see how different nations are positioned with their export compared with the global situation in figure 1.13.

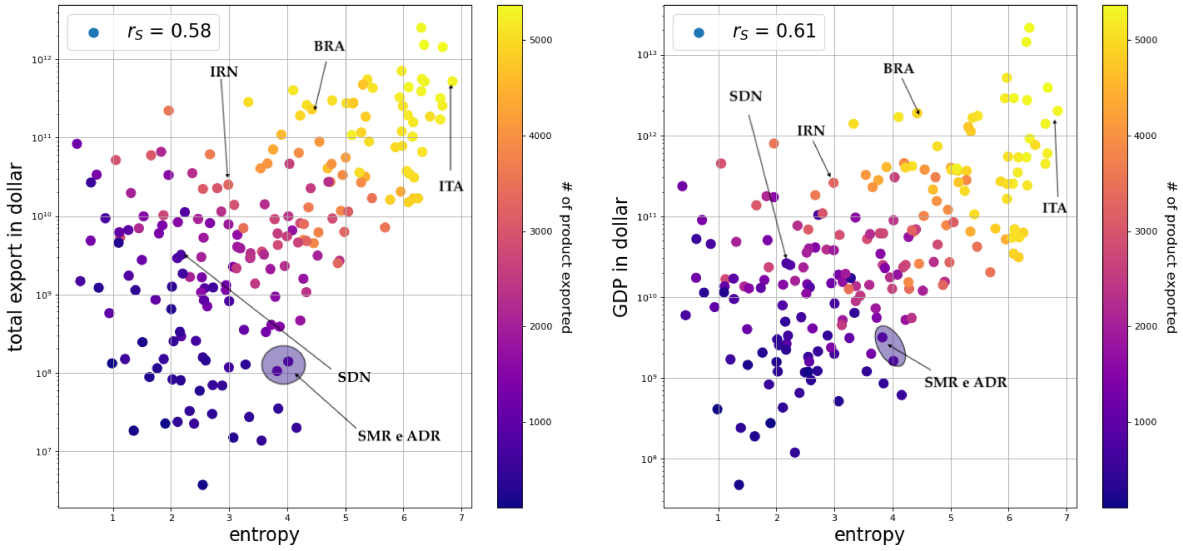


Figure. 1.10: Entropy correlation with extensive monetary measures

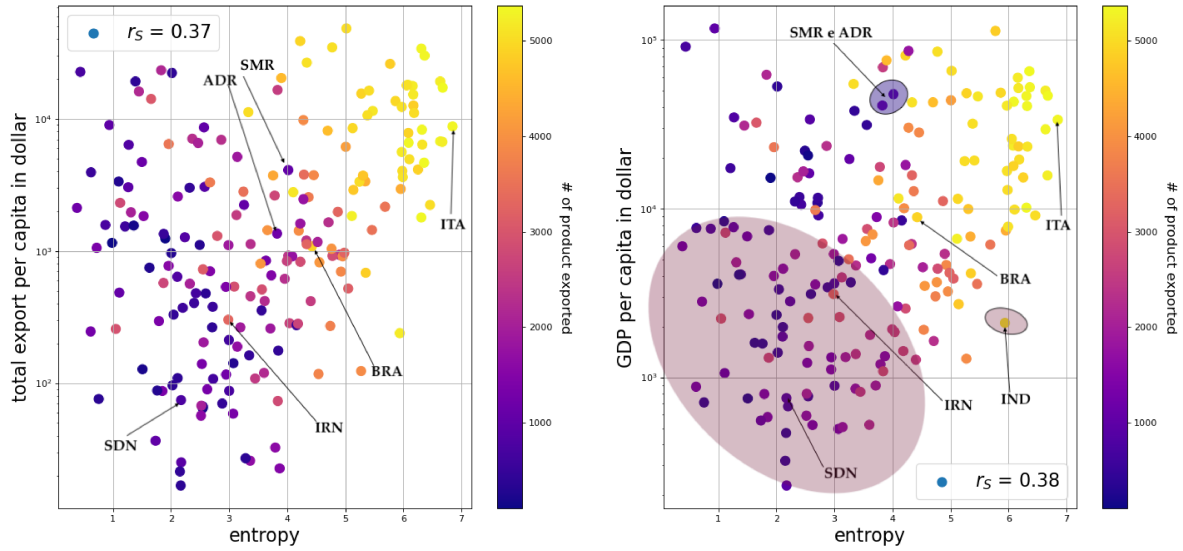


Figure. 1.11: Entropy correlation with intensive monetary measures

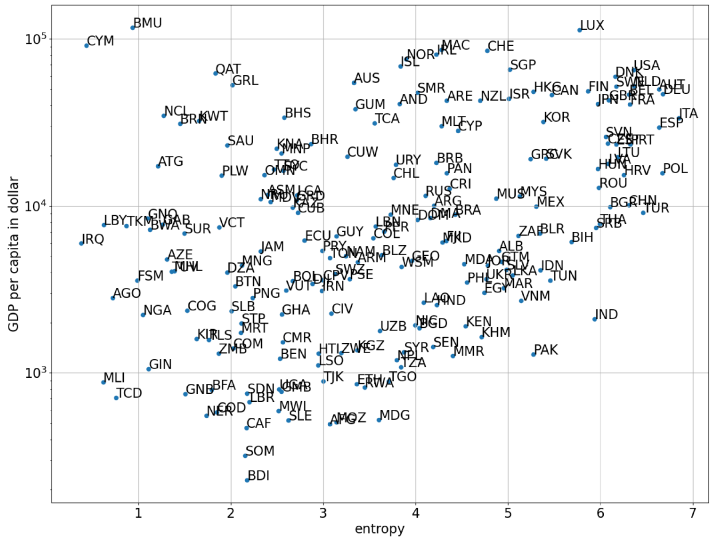


Figure. 1.12: Entropy-GDPpc scatter plot for the year 2019 with HS edition of 2017. On each point is depicted the iso 3 codes of each nation.

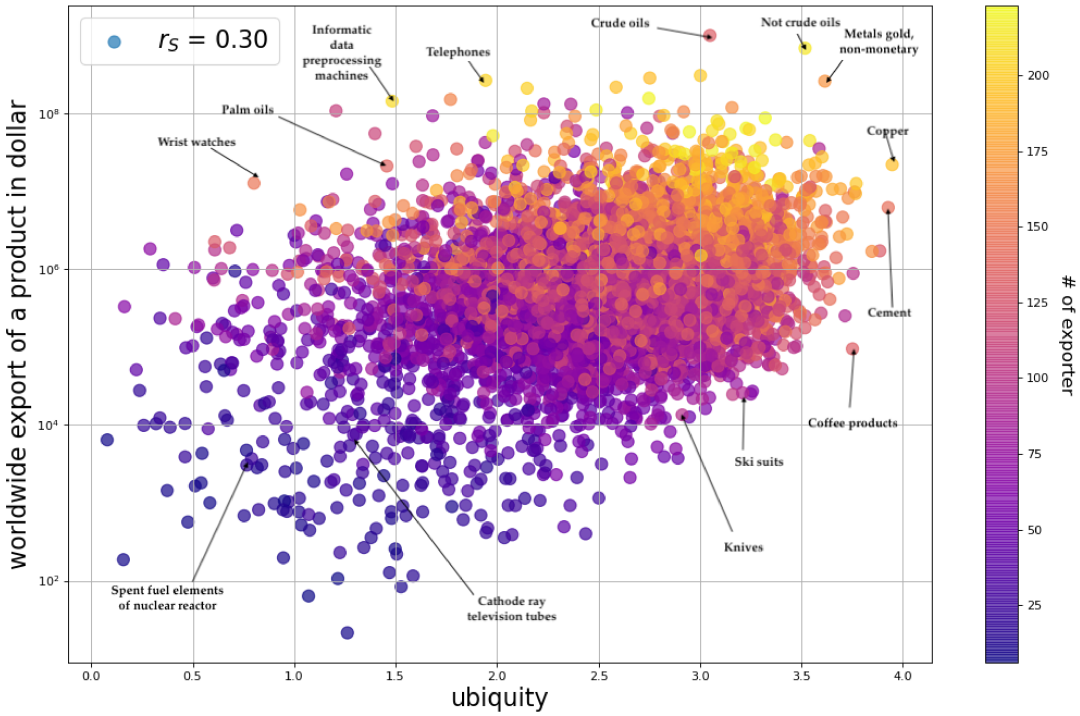


Figure. 1.13: Scatter plot of the product export - ubiquity

1.6 Coarse-Grained Properties

In the other complexity measures, aggregation at 4 or 2 digit level of the Harmonized System is performed in the bi-adjacency matrix, so if P is a macro category of products, an aggregated bi-adjacency matrix can be constructed as

$$X_{cP} = \sum_{p \in P} X_{cp} \tag{1.23}$$

after this step, Fitness or ECI can be computed but losing the finer information. The entropy algorithm allows leveraging a special summation rule when we cluster over products, overlooking this step entirely. Instead of aggregating products in the bi-adjacency matrix, we can cluster them directly in the weighted shares normalized solutions of the entropy algorithm.

$$\xi_{cP} = \sum_{p \in P} \xi_{cp} \tag{1.24}$$

This new coarse-grained share is already normalized

$$\sum_{P} \xi_{cP} = \sum_{P} \sum_{p \in P} \xi_{cp} = \sum_{p} \xi_{cp} = 1 \tag{1.25}$$

A similar approach can be exploited in the weighted shares normalized concerning worldwide product export

$$\zeta_{cP} = \sum_{p \in P} \zeta_{cp} \tag{1.26}$$

However, this new coarse-grained share is no more normalized. A new renormalization is needed

$$\zeta_{cP} = \frac{\sum_{p \in P} \zeta_{cp}}{\sum_{c} \sum_{p \in P} \zeta_{cp}} \tag{1.27}$$

This clearly works also for country aggregation, or both product and country aggregation. We summarize the procedures:

$$\begin{cases} \xi_{cP} = \sum_{p \in P} \xi_{cp} \\ \\ \zeta_{cP} = \dfrac{\sum_{p \in P} \zeta_{cp}}{\sum_{c} \sum_{p \in P} \zeta_{cp}} \end{cases} \qquad \text{product aggregation} \tag{1.28}$$

$$\tag{1.29}$$

$$\begin{cases} \xi_{Cp} = \dfrac{\sum_{c \in C} \xi_{cp}}{\sum_{p} \sum_{c \in C} \xi_{cp}} \\ \\ \zeta_{Cp} = \sum_{c \in C} \zeta_{cp} \end{cases} \qquad \text{country aggregation} \tag{1.30}$$

$$\tag{1.31}$$

$$\begin{cases} \xi_{CP} = \dfrac{\sum_{c,p \in C,P} \xi_{cp}}{\sum_{P} \sum_{c,p \in C,P} \xi_{cp}} \\ \\ \zeta_{CP} = \dfrac{\sum_{c,p \in C,P} \zeta_{cp}}{\sum_{C} \sum_{c,p \in C,P} \zeta_{cp}} \end{cases} \qquad \text{country and product aggregation} \tag{1.32}$$

At first sight, we immediately get an improvement as the aggregation is performed a posteriori on the weighted shares normalized matrix, solutions of the iterative scheme that uses the not aggregated information.
In the following, we will analyze products aggregation from 6 to 4 digits for the year 2019 and HS edition 2017.

1.6.1 Coarse-grained Shannon entropy

From the coarse-grained weighted shares normalized we can compute a coarse-grained entropy

$$H_c^{CG} = -\sum_P \xi_{cP} \log(\xi_{cP}) \tag{1.33}$$

This entropy is always less than or equal to H_c.

Proof $H_c \geq H_c^{CG}$

$$\begin{aligned} H_c - H_c^{CG} &= -\sum_p \xi_{cp} \log(\xi_{cp}) + \sum_P \xi_{cP} \log(\xi_{cP}) \\ &= -\sum_p \xi_{cp} \log(\xi_{cp}) + \sum_P \sum_{p \in P} \xi_{cp} \log(\xi_{cP}) \\ &= -\sum_p \xi_{cp} \big[\log(\xi_{cp}) - \log(\xi_{cP}) \big] \\ &= -\sum_p \xi_{cp} \log\left(\frac{\xi_{cp}}{\xi_{cP}}\right) \\ &= \sum_P \xi_{cP} \left[-\sum_{p \in P} \frac{\xi_{cp}}{\xi_{cP}} \log\left(\frac{\xi_{cp}}{\xi_{cP}}\right) \right] = \sum_P \xi_{cP} H_{cP} \end{aligned}$$

At the end we have defined a new quantity H_{cP}, which is actually an entropy as the argument of the Shannon entropy sum to one

$$H_{cP} = -\sum_{p \in P} \frac{\xi_{cp}}{\xi_{cP}} \log\left(\frac{\xi_{cp}}{\xi_{cP}}\right) \tag{1.34}$$

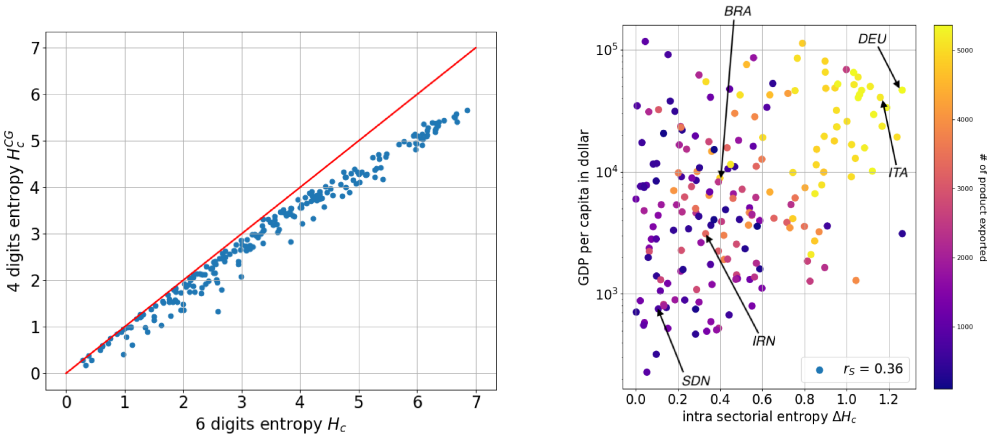
This entropy, called **intra-sectorial entropy**(1.34) measures the diversification of a country into a P macro-category of products. Therefore, using the fact that $H_{cP} \geq 0$ we end up with $H_c \geq H_c^{CG}$. The equality holds in the case where for every coarse-grained category P there is one and only one fine-grained category p.

The difference $H_c - H_c^{CG}$, called **inter-sectorial entropy**(1.35), measures the gain in diversification that a country obtains, focusing on diffusing its productive system into a more specialized micro category of goods.

$$\Delta H_c = H_c - H_c^{CG} = \sum_P \xi_{cP} H_{cP} \tag{1.35}$$

In figure 1.14 we observe the inter-sectorial entropy, computed using aggregation from 6 to 4 digits of products of the Harmonized System, against GDPpc. Wealthy countries show an extreme articulated structure of the export shares (we find Germany at the first position), while

countries in the poverty region have a scarce inner organization in their economic structure. We also find oil-dependent countries, such as Saudi Arabia, with a low intra-sectorial entropy, indicating that their export shares are concentrated in few sectors. The other country that has a very high inner economic organization is Vanuatu, a small nation in the Pacific Ocean, indicating that, besides its very low diversification in export (it exports 372 products, on a total of 5400, and has an entropy of $H_c = 2.6$, in contrast with Germany that has $H_c = 6.8$), it has a very diversified little economy. This example shows that an inter-sectorial analysis has to be coupled with a sectorial one to get meaningful information.



(a) 6 digit entropy vs. 4 digit entropy

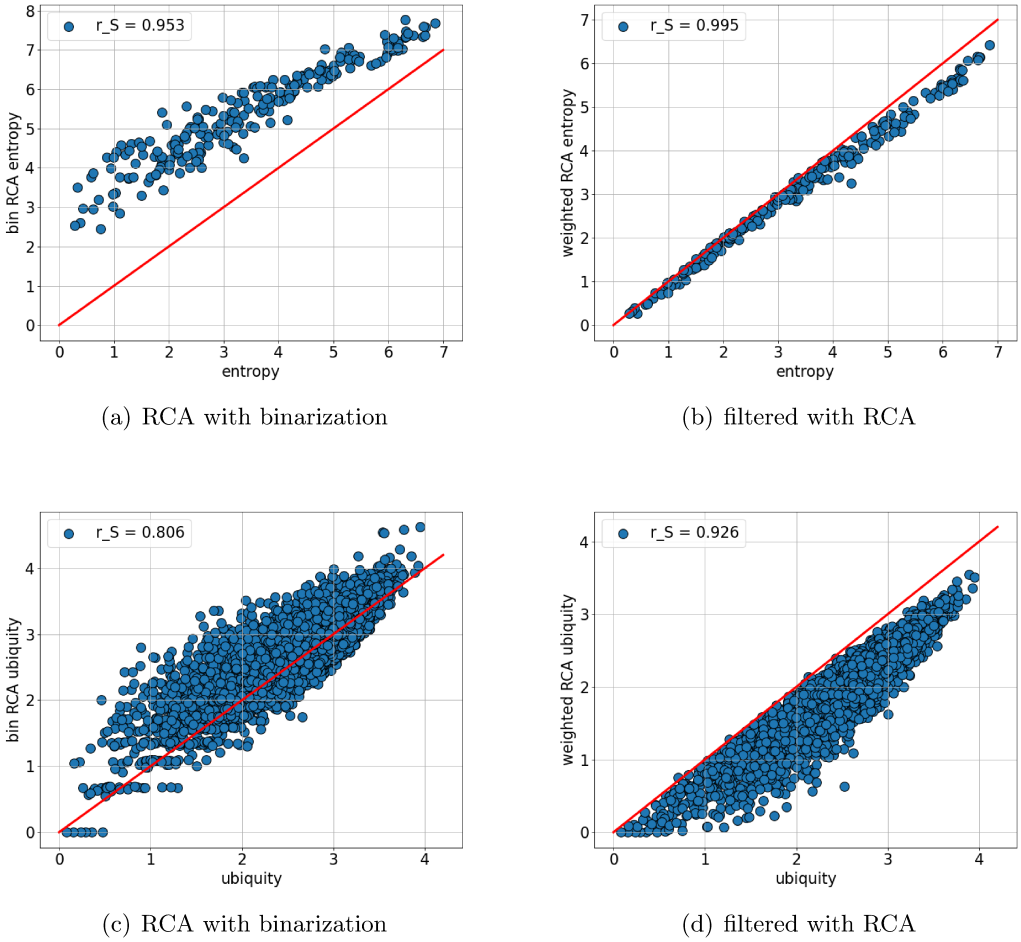
(b) inter-sectorial entropy

Figure. 1.14: In figure (a) is evident the inequality $H_c > H_c^{CG}$. Developed countries have a high inner organization of their economic structure, while poor countries show a scarce inner organization.

1.7 Difference with RCA, Binarization and HS Edition

The stability of the entropy algorithm allows tackling the problem of ranking nodes in a bipartite graph in many different ways. In the initial part of the chapter, we have defined two different ways to construct the bi-adjacency matrix(1.2), using the revealed comparative advantage criterion (RCA), common in most of the works in economic complexity. In this section, we want to discuss the role of RCA in the entropy algorithm framework.
From the bi-adjacency matrix $\tilde{X}_{cp}$, we can calculate entropies indicating the level of diversification of nations as if they exported only goods over which they have a revealed comparative advantage; we call these indicators *filtered RCA entropies*. Using the binarized bi-adjacency matrix, we eliminate the monetary information about products (to be precise, it is still into the RCA criterion) and their different weights into the basket of a nation; it is like considering it an urn full of equal balls (products). The relative weights of the balls into the urn are given by the entropy algorithm, giving heterogeneity to the baskets of countries. In this sense, the *binarized RCA entropy* is not an indicator of diversification of the basket of export but a complex count of products on which is revealed comparative advantage. This last approach is more similar to Fitness and ECI in concept.

In figure 1.15, we observe the variation of our complexity measure using the two different RCA criteria. As expected with the RCA criterion, with no binarization (figure (b)), we get smaller entropic values as the matrix is the same with fewer non-zero elements; therefore, the probability distributions of countries and products move away from the uniform distribution. This is a



(a) RCA with binarization

(b) filtered with RCA

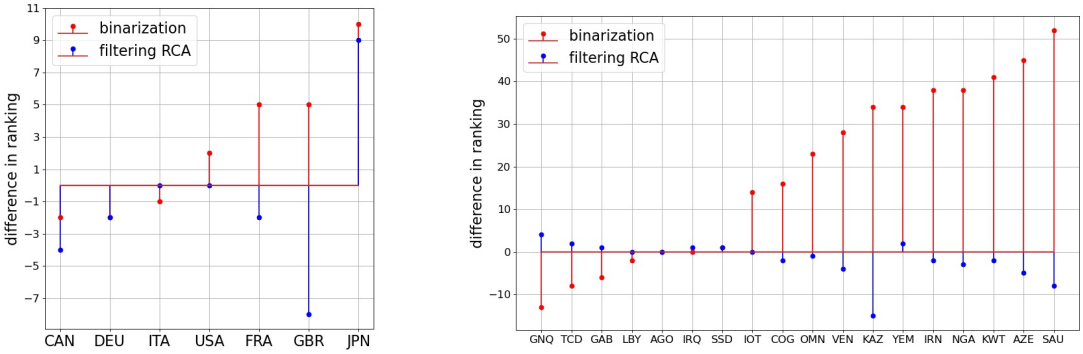
(c) RCA with binarization

(d) filtered with RCA

Figure. 1.15: In figure (a) we observed the entropy measure obtained using the binarization criteria M_{cp} versus the one obtained with the full dataset, the correlation among ranks is very high ($r_S = 0.95$). In figure (b) the entropy obtained with the matrix $\tilde{X}_{cp}$ is plotted against the full information entropy, it correlates very well ($r_S = 0.99$). In figure (c) is depicted the binarized ubiquity with RCA against the full information ubiquity ($r_S = 0.81$), the correlation is still high. In the last figure (d) we have the filtered RCA ubiquity against the full information one ($r_S = 0.93$)

consequence of the fact that for coarse-grained entropy, as the filtered weighted matrix is, is valid this inequality $H > H^{CG}$. As visible in figure (d) and (b), a few dots can been found slightly above the red line. In fact, this relation is non more exact after the iteration process, as different initial bi-adjacency matrices give different weights to the normalized weighted shares. Regarding the entropies computed with binarization criteria the loss of information is evident, as the elements tend to be concentrated in areas with higher entropy; still a high correlation in ranks remains.
The difference among ubiquity measures is more evident: with a RCA filter we still observe the relation $H > H^{CG}$ and an high correlation ($r_S = 0.93$), but with binarization the correlation gets smaller ($r_S = 0.81$).

Regarding the different ranking for nations, we have decided to consider G7 countries Figure 1.16(a) (for the most developed), top oil-dependent countries Figure 1.16(b) (with a share of oil export in their basket higher than 0.5), G20 countries Figure 1.17(a) (for middle developed nations), and also countries with a small population, less than 1 million. Analyzing G7



(a) G7 countries

(b) Oil dependent countries

Figure. 1.16: In figure (a) we have the countries that are into the G7 organization, the plot describes the differences in ranking due to two different procedures of data pre-processing with RCA. In figure (b) we have the most oil-dependent countries (shares of oil export more than 0.5).

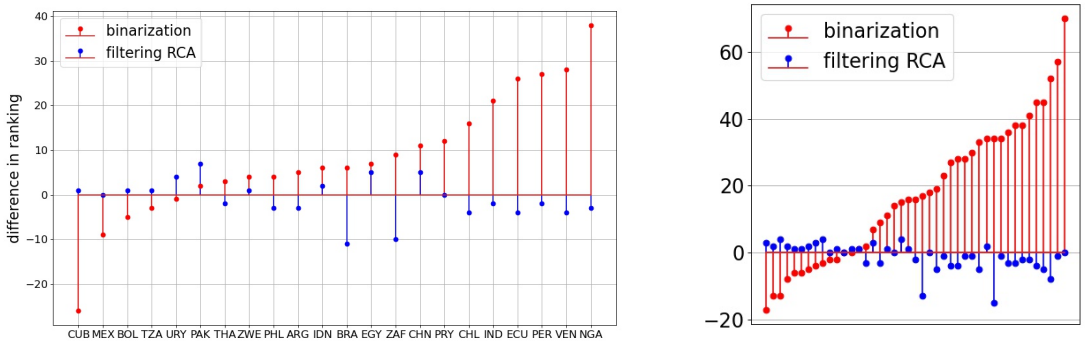
countries, we observe that Japan has the most significant positive difference, indicating a solid revealed comparative advantage on many products. Interestingly, the binarization criterion for oil-dependent countries tends to give higher ranks, but if we filter with RCA, this does not happen. With binarization, we lose the information about the relevance of some products in a country's basket, giving them higher ranks at the end.
For G20 countries, apart from the last two countries (Nigeria and Venezuela) that are also oil-dependent ones, we find Perù (PER) and Chile (CHL) that both have a fraction of the 26% of copper exportation, and Ecuador (ECU) that is still oil-dependent (32% of oil in the basket of export). These considerations brought us to consider countries that are one-product dependent Figure 1.17(b) : a product that constitutes more than half of their basket of export. It is visible that the binarization criterion tends to give higher ranks to these countries. As expected, crude petroleum constitutes the most common and significant product for these countries. However, we can also find copper (Zimbabwe), Gold (Uganda, Suriname, Somalia, Nigeria, Mali), petroleum gases (Turkmenistan), tobacco (Malawi), aluminum oxide (Jamaica), mollusks (Falkland islands), and diamonds (Botswana).
If we study the ranks of nations with less than one million population, we find that the binarization criterion tends to lower those countries' ranks. Indeed, it is remarkable that the binarized entropy correlates better with population ($r_S = 0.43$) than full information entropy (as pointed out in the results section $r_S = 0.35$) and filtered RCA entropy ($r_S = 0.34$). Concerning the pre-processing data phase with filtering with RCA, we have not noted significant differences, thanks also to the high correlation in ranks with full information entropy ($r_S = 0.99$). We summarize these considerations

a. RCA binarization tends to give higher ranks to one-product-dependent countries, as it does not consider the weight that such export has on their basket.

b. An higher correlation with population has been found using RCA binarization, ($r_S = 0.43$) against ($r_S = 0.34$).

c. The distribution of the entropies computed using RCA binarization is more concentrated than the other distributions (see 1.19).

Further studies on the role of RCA can be done by looking at how the basket of a country changes



(a) G20 countries

(b) product dependent countries

Figure. 1.17: In figure (a) are depicted the differences in ranking for the G20 countries. In plot (b) we have the countries that are product dependent, that have one product that constitutes half of their basket of export.

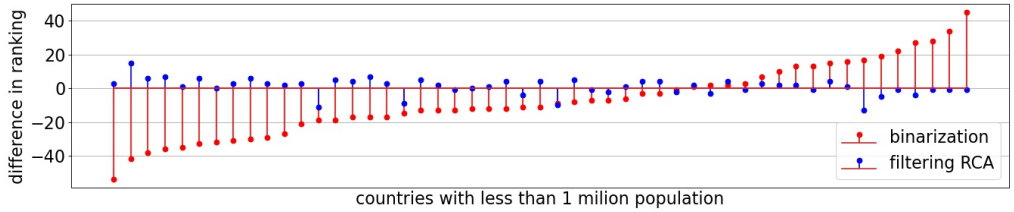
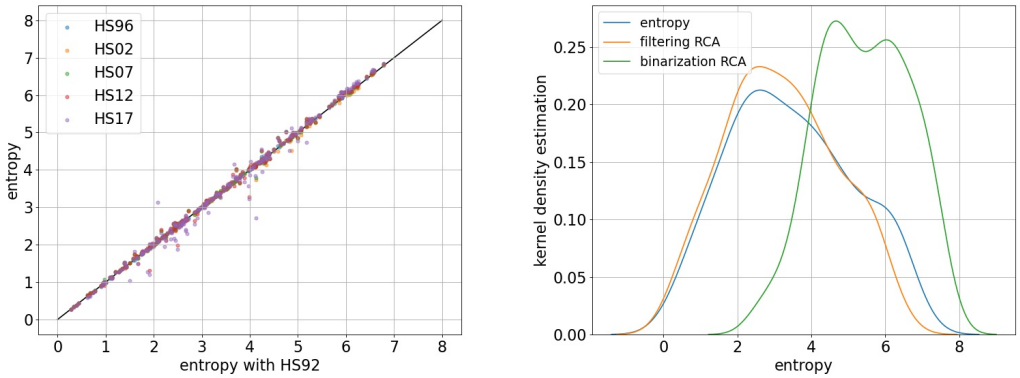


Figure. 1.18: Countries with less than 1 milion population

when we apply the filter. The top result that emerges from figure 1.21 is that RCA filter removes all the products that have a low contribution in the basket. The red line in the colorbar of the right panels shows the minimum value of the shares of a product after RCA filtering, while the black lines are the first, second, and third quantile of the distribution of the shares of the products that have no revealed comparative advantage. For under and lower-middle developed countries, these quantiles show that the RCA filter removes all the products that count low in their basket. For instance, we can consider Sudan, which has the red line over the third quantile, indicating that the RCA filter erased over a 75% of products indistinctly below the red line. However, especially if we want to study the dynamics of these countries' entropies, it seems incorrect to say that they do not have at all the capabilities and the know-how related to those products. The evolution in entropy and ranks of these countries should depend primarily on diversifying and getting more competitive to products that have low importance in their basket. Therefore, it seems somewhat unclear why we should consider an RCA filter in our approach.
Those plots also show the heterogeneity in the basket among nations. The red lines indicate the median of ubiquity and total export for each country, and interestingly, we observe that the more a nation is wealthy, the more its median tends to be the same as that of an ideal world country.
These considerations corroborate our idea of not using RCA criteria within the construction of the bipartite graph and so support the adoption of our complexity measure. An analysis of the dynamics of these different measures will be studied in the next chapter.

Different HS edition BACI provides 6 different editions of the Harmonized System (1992, 1996, 2002, 2007, 2012, 2017), a new edition of 2022 has been recently released, but there are

still no BACI datasets with this classification. An analysis of how the country's ranks change with different editions of the HS has to be made because only for the first edition, we have a wide range of years from 1995 to 2019. In Figure 1.18(a) we can see the correlation of the measures among the different edition of the HS. The correlation with the next edition is very high (HS92-HS96: $r_s = 0.9999$) and it constantly decrease with the new editions (HS92-HS17: $r_S = 0.99$). Still, a very high correlation among ranks remains.



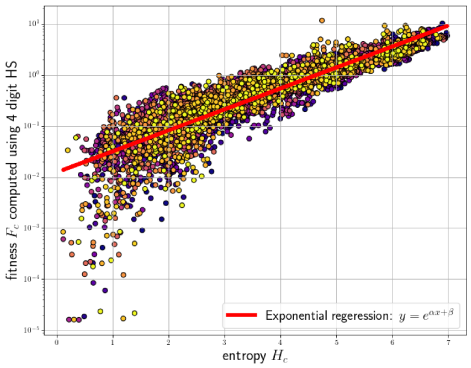
(a) Different HS edition agains HS 2017

(b) pdf of the measures

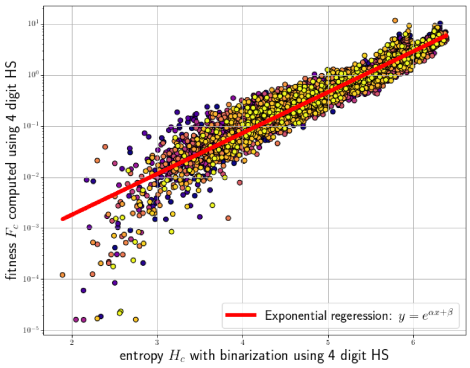
Figure. 1.19: In figure (a) we have a scatter plot of the various entropies that can be computed using different editions of the harmonized system, we have a Spearman coefficient of 0.99 for every edition. In plot (b) we observe the different distribution of entropy, computed using Gaussian kernel density estimation with a bandwidth of 1, the distribution of the binarized entropies is more concentrated than the others.

1.8 Comparison with Fitness

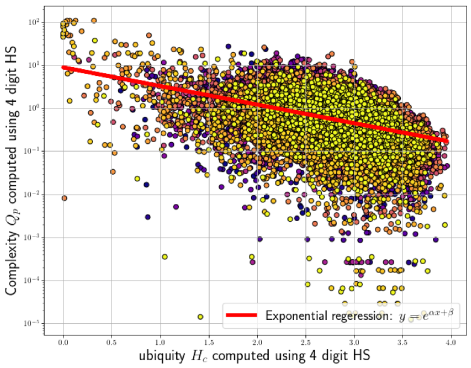
Fitness algorithm is commonly accepted, and a lot of ink has been spilled using it. It is therefore appropriate a direct comparison with this indicator. From 1995 to 2019, we surprisingly obtained an excellent correlation between fitness (computed using 4 digits of the HS92 and binarization procedure) and entropy (computed both with full information and 6 digits and binarization with 4 digits, using HS96), see figure 1.20. Each color of the scatter plot represents a particular year, from 1995 to 2019, with the only exception that we had to remove fitness results of 2018 because, for some reason, that dataset showed very poor convergence using the HS edition of 1992. A few country's measures did not converge using the fitness method, therefore we had to remove the point smaller the 10^{-5} (also for ubiquity) and fitness was taken at 200 iterations while entropy was taken at 100 iteration steps. The Spearman correlation coefficient $r_S = 0.94$ suggests the following relation $F_c = \exp(\alpha H_c + \beta)$ and a fit procedure gave as parameters $\alpha = 0.995 \pm 0.007$ and $\beta = -4.39 \pm 0.03$.
We have computed the ubiquity of a product using the first 4 digits of the HS92 through the coarse-grained procedure. The relation with the complexity measure is shown in Figure 1.20 (c), a low absolute correlation is visible as the Spearman correlation coefficient is $r_S = -0.43$ suggesting that complexity and ubiquity catch different aspects about products. The negative sign of the coefficient validates the opposite interpretation we can give to those measures: more complexity is given to less ubiquitous products.



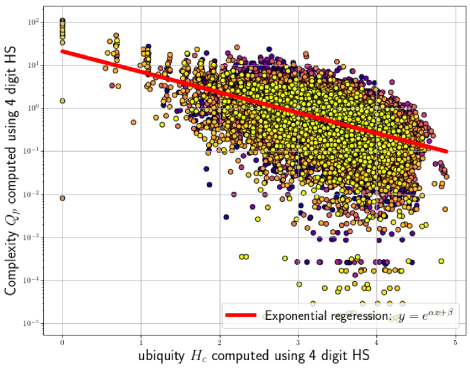
(a) fintess (4 digit)-entropy (6 digit)



(b) fintess (4 digit)-binarized entropy (4 digit)



(c) complexity (4 digit)-ubiquity (4 digit)



(d) complexity (4 digit)-binarized ubiquity (4 digit)

Figure. 1.20: Each color point represents a particular year from 1995 to 2019, the dataset is BACI with HS edition 1992. In fig (a) and (b) it is visible a strong correlation between entropy and fitness (Spearman correlator for figure (a): 0.94, while for figure (b): 0.96), the correlation is stronger if we use the same structure of the bipartite network as in figure (b). An exponential fit suggests an exponential map between the two measures. In fig (c) and (d) a smaller correlation is visible between the complexity and ubiquity (Spearman correlator for figure (c): -0.43, while for figure (d): -0.54), the negative value of the Spearman coefficient validates the opposite meaning between complexity and ubiquity of a product. We have the following parameters for the fits. An important note: we had to remove fitness for the year 2018 for convergence problems.
(a): $\alpha = 0.995 \pm 0.007$ $\beta = -4.39 \pm 0.03$
(b): $\alpha = 1.841 \pm 0.009$ $\beta = -9.98 \pm 0.04$
(c): $\alpha = -0.99 \pm 0.01$ $\beta = 2.18 \pm 0.04$
(d): $\alpha = -1.09 \pm 0.01$ $\beta = 3.03 \pm 0.04$

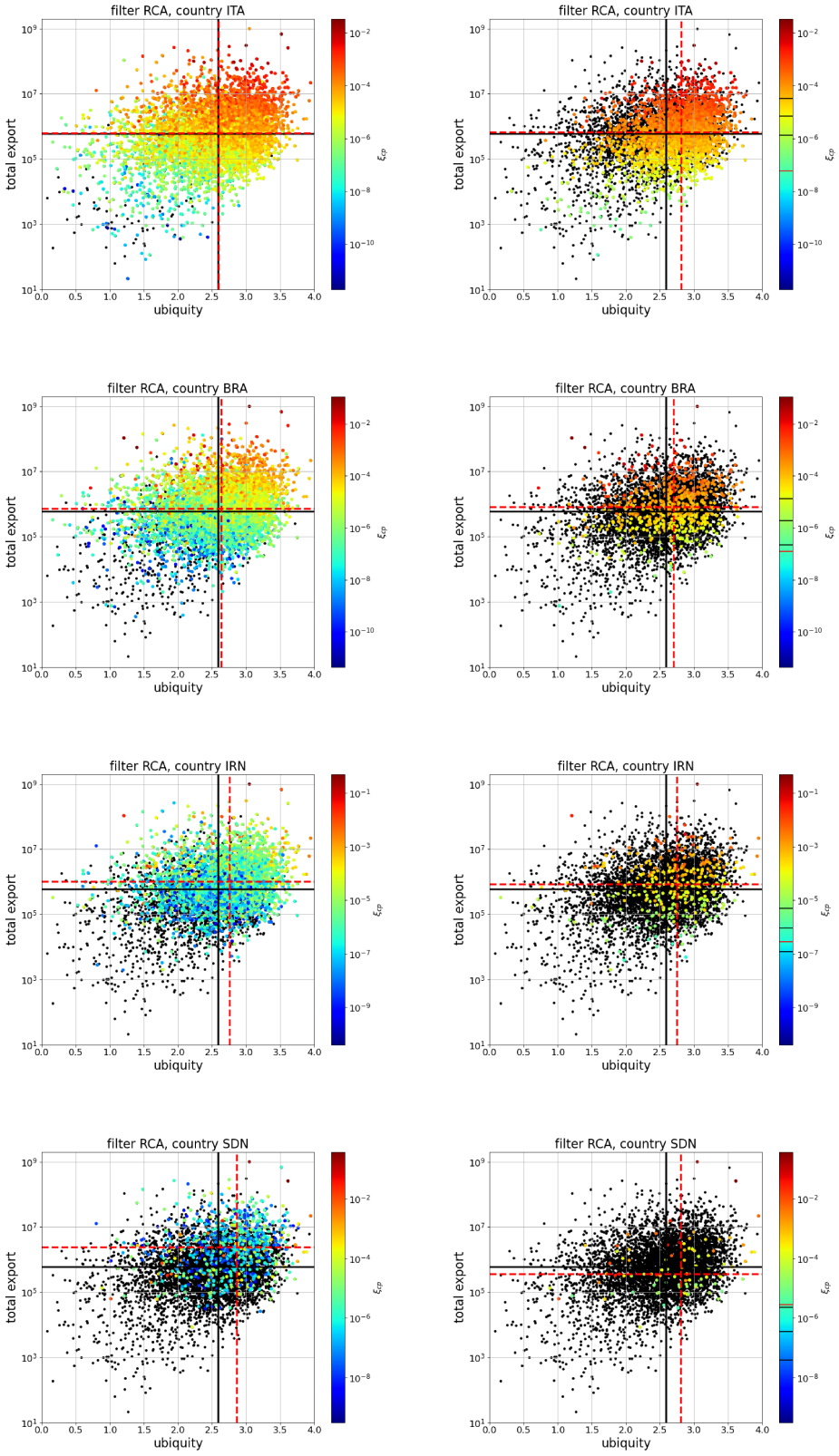


Figure. 1.21: Basket composition of countries, black dots represent all products at 6 digits of HS17. We have the basket distribution in the left panels, represented using a colorbar, while in the right panels, we show the same distribution after an RCA filter. High, upper-middle, lower-middle, and low developed countries are depicted. Black lines are the median for the entire dataframe of products, while red lines represent the median of the actual composition of the basket. Black horizontal lines in the right colorbar are the first, second and third quantile of the distribution of the removed shares due to filtering, while red lines indicate the minor shares after the filter.

Dynamics in the entropy-monetary plane

In the previous chapter, we studied how a complexity measure, based on Shannon entropy, of a complex bipartite network can be constructed from BACI datasets that gather and process information about worldwide export. The analysis was done only for the most recent year available (which is 2019) and using the most recent edition of the Harmonized System (HS), the 2017 edition that at 6 digits level accounts for almost 5400 products. However, within this edition of the HS data are available only for three years (2017, 2018, and 2019); therefore, it would not be an excellent choice for the study of the dynamics of the complexity measure. The richest dataset offered by BACI is the one with the first edition of the HS (1992), which contains fewer products (about 5000), decreasing over the years, as a consequence of the fact that some products that had been listed in 1992 were no more exported in the past few years. Within this edition, we have access to the larger time window from 1995 to 2019. Other editions are available in BACI (see [52]) but they cover smaller time windows.
We computed entropy measures from 1995 to 2019, the evolution of the ranks is visible in Figure 2.1, on which we can notice that countries with very high entropy tend to remain in those positions. Remarkably, Italy turns out to have the highest entropy all over the time window, indicating a robust export diversification.

The biggest jumps towards high ranks were made by Poland (from 17 to 2), Spain (from 16 to 3), Turkey (from 20 to 6), and Portugal (from 23 to 8). In contrast, the biggest jumps in the opposite direction were made by Hong Kong (from 13 to 37), the Czech Republic (from 3 to 20), China (from 4 to 13), and Switzerland (from 15 to 56), the most significant downward jump was made by Australia (from 40 to 100).
In contrast, countries with the lowest entropic values tend not to change their positions. For instance, we can consider Iraq, which has the lowest entropy in all the time windows (except in 1995). Its constant low position is a consequence of the fact that its export is mainly characterized by a single product: oil. In the area with low entropy (right panel), there are a lot of upward movements, which is a sign that poor diversified countries tend to diversify their export. Big upward jumps were made by Samoa (from 164 to 75), United Arab Emirates (from 121 to 64), Cambodia (from 109 to 54), and Iran (from 160 to 108).
In the middle region, there is much movement in both directions, and an example would be Egypt (from 76 to 38) and Uruguay (from 51 to 87).

The dynamics in the entropic dimension seems to be rather chaotic, so to better understand it is necessary to add a new dimension. A first idea could be to couple entropy and export per capita. We have seen that entropy correlates less with per capita monetary indicators, suggesting that the two measures embed different information and could be coupled to observe dynamical patterns Figure 2.2. We will call this graph *entropy-monetary plane*.

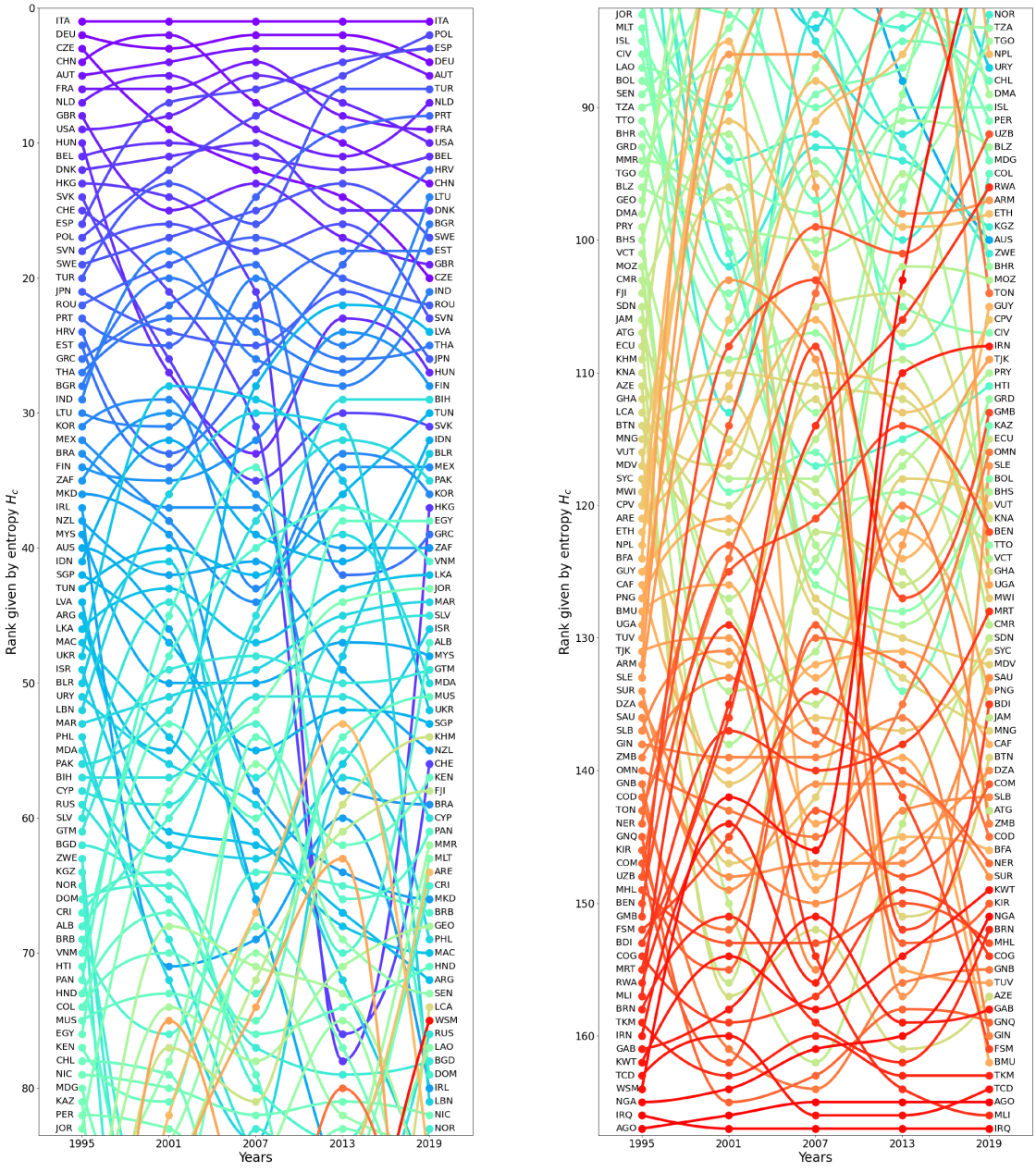


Figure. 2.1: Evolution of the ranks of the countries given by the entropic complexity measure for the years 1995, 2001, 2007, 2013, 2019. Entropies are computed using the dataset BACI with Harmonized System edition 1992.

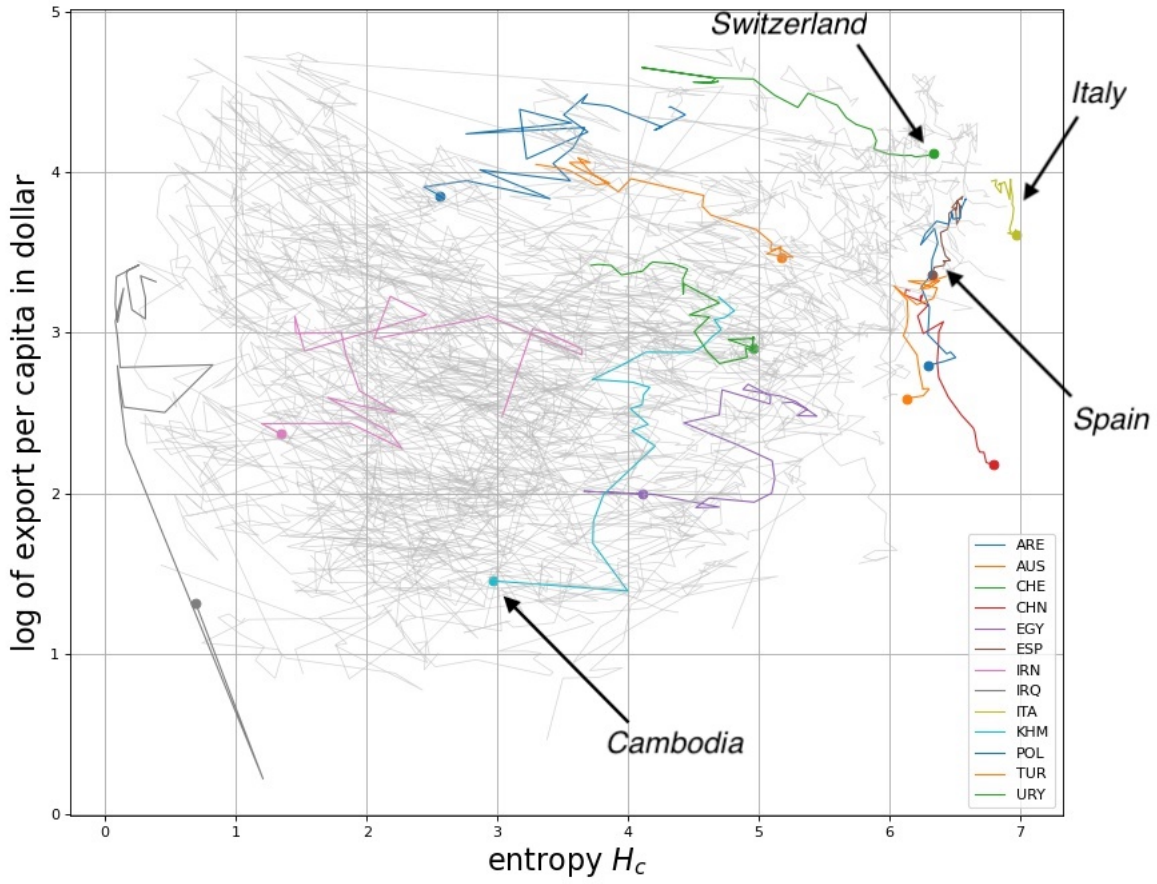


Figure. 2.2: Dynamics in the entropy - export per capita plane. Dots represent the initial point (1995).

This very first graph already gives some insight. Let us take, for instance, Switzerland (CHE) and Spain (ESP). Initially, they have a similar entropy, but the first evolves towards smaller entropy while the other does the opposite. The difference is in their position in the plane: Switzerland already has a considerable export per capita while Spain does not. Moreover, countries with similar conditions to Spain (see Poland and Turkey) appear to evolve similarly in the plane. It is also visible that middle development counties have a more complex dynamics than developed countries; for example, Italy (ITA) has a more stable dynamics and growth than Cambodia (KHM).

2.1 Coarse-Graining the Plane

Following the procedure introduced in [35] for the Fitness method, we can investigate the graph in Figure 2.2 better by dividing the plane into boxes. For each one we will compute a trend as an average of all the velocities in that box. In this way, we can extract a flow structure from the graph and investigate a coarse-grained dynamics.
For each box b, we collected the set of points that lay in that box $\{\mathbf{x}_{c,t}\}_b$, where $\mathbf{x}$ is a point in

the entropy-income plane $\mathbf{x} = (H_c, Y)$, H_c is the entropy and Y is a general monetary indicator of a nation (in the considered graph Y indicates the logarithm of the total export), while c indicates a country and t a particular year.
Each point in the coarse-grained box has a one-year evolution $\{\mathbf{x}_{c,t+1}\}$ that could be anywhere in the plane. The velocity, or trend, can be calculated as a one-year displacement, regardless of whether or not the evolution is inside the box.

$$\{\mathbf{x}_{c,t}\}_b \longrightarrow \{\mathbf{v}_{c,t}\}_b = \{\mathbf{x}_{c,t+1} - \mathbf{x}_{c,t}\} \qquad \mathbf{x}_{c,t} \text{ into the } b \text{ coarse-grained box} \tag{2.1}$$

Whit this construction we get a set of velocities $\{\mathbf{v}_{c,t}\}_b$ for each coarse-grained box b.

Central tendency trend The aim is to compute an unique trend, or central tendency trend, from this set in order to assign an arrow to each box. To do this we can extract an average from $\{\mathbf{v}_{c,t}\}_b$ as done in [35].

$$\langle \mathbf{v} \rangle_b = \frac{\sum_{(c,t)\in b} \mathbf{v}_{c,t}}{n_b} \tag{2.2}$$

Where the sum is over the set $\{\mathbf{v}_{c,t}\}_b$ and n_b is the cardinality of the set, hence the number of velocities into the coarse-grained box b.
However, this approach could be uninformative about the middle points of a country's time series. For instance, let us consider the case of a country's time series $\{\mathbf{x}_{c,t}\}_{t=T_0,\dots,T}$ (now we fix c and consider the series in t from the initial time T_0 to the final time T) remaining into a coarse-grained box b. If we compute a simple average of its one-year velocities, we will get a normalized displacement from the initial point $\mathbf{x}_{c,T_0}$ to the final point $\mathbf{x}_{c,T}$, loosing the information about the middle evolution. Indicating as $\langle \mathbf{v}_c \rangle_b$ the average one-year trend of the country's time series, and $n_b(c)$ the number of points of the time series into the coarse-grained box b, we have indeed

$$\langle \mathbf{v}_c \rangle_b = \frac{\sum_{t\in b}(\mathbf{x}_{c,t+1} - \mathbf{x}_{c,t})}{n_b(c)} \tag{2.3}$$

$$= \frac{(\mathbf{x}_{c,T} - \mathbf{x}_{c,T-1}) + (\mathbf{x}_{c,T-1} - \mathbf{x}_{c,T-2}) + \cdots + (\mathbf{x}_{c,T_0+1} - \mathbf{x}_{c,T_0})}{n_b(c)} \tag{2.4}$$

$$= \frac{\mathbf{x}_{c,T} - \mathbf{x}_{c,T_0}}{n_b(c)} \tag{2.5}$$

Moreover, the set of velocities $\{\mathbf{v}_{c,t}\}_b$ is affected by outliers that can arise due to particular economical situations, such as economic crisis. The arithmetic mean in (2.2) is greatly influenced by outliers, making the indicator $\langle \mathbf{v} \rangle_b$ mostly dependent on large displacements. We decide to replace this average with the median, which is a robust measure of central tendency and so better represents a "typical" trend. In addition, with this indicator we no longer have the problem of the uninformative middle evolution.

$$\langle \mathbf{v} \rangle_b = \text{median}(\{\mathbf{v}_{c,t}\}_b) \tag{2.6}$$

Collinearity Measure The central tendency trend of a coarse-grained box give information about the most likely one year evolution of points inside it. However, this information is not sufficient to properly catch the chaos embedded in the coarse-grained box; it gives only the overall trend and does not say anything about the distribution of the velocities that it contains. A possible measure of chaos into a coarse-grained box b consists in taking the trace of the covariance matrix computed using all the velocities $\{\mathbf{v}_{c,t}\}_b$, normalized with the area of the box

Σ_b. Since the trace of a covariance matrix is the sum of the variances of the coordinates of the entropy-monetary plane, we indicate this measure of chaos as σ_b^2.

$$\sigma_b^2 = \frac{\mathrm{Tr}\big(\mathrm{Cov}(\{\mathbf{v}_{c,t}\}_b\big)}{\Sigma_b} \tag{2.7}$$

The higher is σ_b^2 the more chaotic dynamics we get in the box, while the lower it is the more collinear are the velocities in the box. In this sense the chaos' measure σ_b^2 is related to a measure of collinearity $\mathcal{K}_b$ of the velocity set. In the following we will use both term to indicate chaos or collinearity.

$$\sigma_b^2 = \frac{1}{\mathcal{K}_b} \tag{2.8}$$

The σ_b^2 represents a scalar field in the coarse-grained entropy-monetary plane that can be pictured as a color in the arrows representing the central tendency trend (2.6). In figure 2.3 we performed a coarse-graining on the entropy-log export plane using 256 boxes of equal area, the arrows are computed using (2.6) while the color indicates the chaos measure (2.7).
A clear separation in colors, therefore in collinearity, and in trend dynamics can be seen: high entropic countries have high collinearity while low entropic countries have the opposite. The meaning is that countries with high entropy evolve similarly, while the others do not.

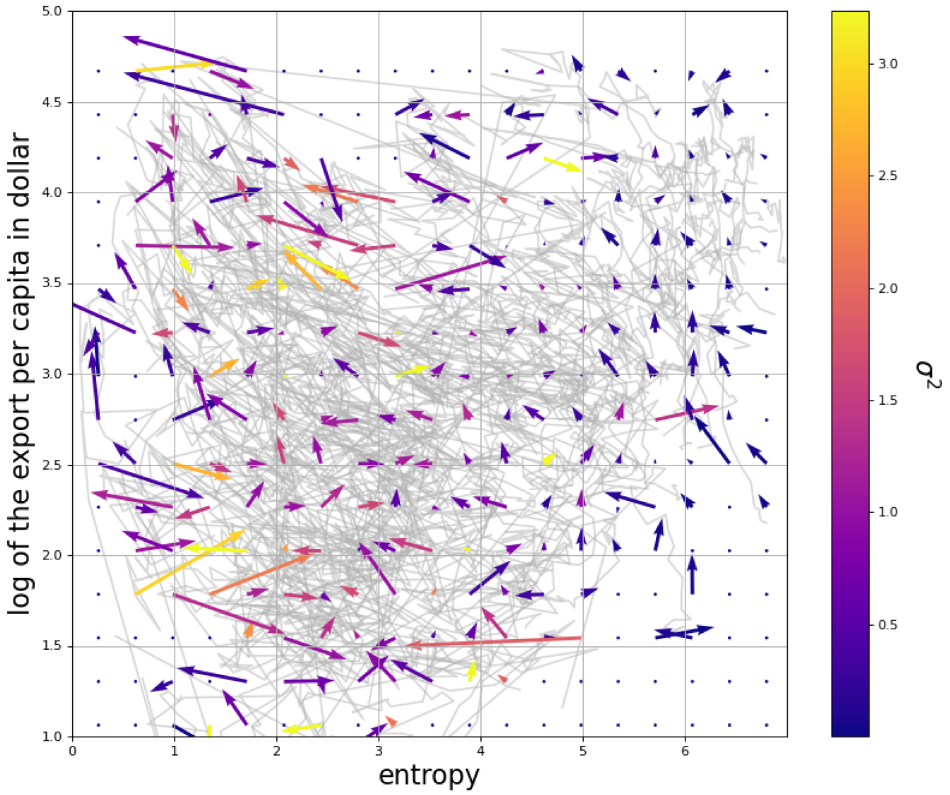


Figure. 2.3: Flows in the entropy-export per capita plane

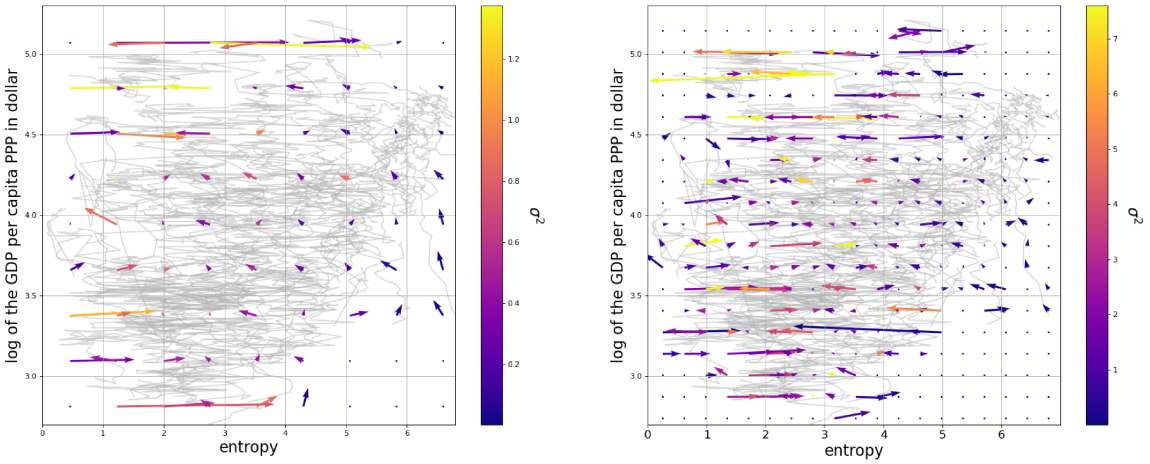


Figure. 2.4: Flow in the entropy - GDPpcPPP plane, the right panel has a fine grained double than the left panel

We also notice that the arrows are more collinear in the high entropy region, indicating that there is also global high collinearity for regions with high entropy. On the other hand, low entropy regions show high variability in the arrows, suggesting that the in-box low collinearity is also present at bigger coarse-grained scales.

The problem of outliers is present also in the chaos measure as they were found in the distribution of $\{\sigma_b^2\}$; this brought some problems in picturing the arrow's color, as the colorbar was entirely overwhelmed by a few outliers. To solve this and properly picture the gradient in color among low and high entropy regions, we removed these outliers by projecting them on a threshold value computed as the quantile of the distribution of the chaos measures corresponding to 0.95: we project 5% of the highest values. In the following we will omit the pedix b that indicates that the chaos measure is related to the b box, hence $\sigma_b^2 = \sigma^2$.

A new monetary measure Despite having extracted some information with this dynamics, we have to highlight some problems in this analysis. Firstly, BACI datasets give export data in current dollars; this means that the export per capita we have computed is affected by inflation. Moreover, inflation is different for each country, and to properly compare monetary values within countries is necessary to consider export (or GDP) weighted by the *Purchased Power Parity*[1]. If PPP solves the problem of comparing monetary values within different countries, it remains the problem of the dynamics of those values that can be affected by world inflation. Therefore, current dollars have to be replaced by *constant* dollars. Fortunately, World Bank provides a dataset of GDP per capita weighted by PPP in constant 2017 international dollar. From now on, we will use this monetary measure in the plane, indicating it as GDPpcPPP (gross domestic product per capita in purchased power parity). All the measure in GDPpcPPP will be reported in a logarithmic scale, so often we will omit this indication.

We show the results with two different choice for the number of coarse-grained boxes for the entropy - GDPpcPPP plane in Figure 2.4. Clearly, with a smaller number of boxes the dynamics

[1]Purchased Power Parity (PPP) is the measurement of prices in different countries that uses the prices of specific goods to compare the absolute purchasing power of the countries' currencies.

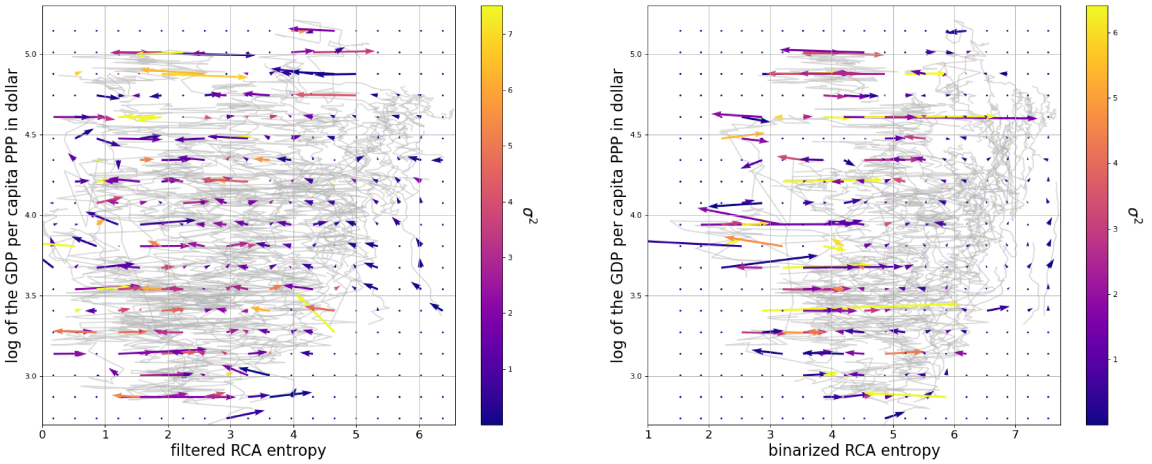


Figure. 2.5: Flow in the entropy - GDPpcPPP plane, left panel is with the binarized version of entropy, while the right one is with the weighted entropy (the one computed using the complete biadjacency matrix X_{cp} filtered with RCA).

seem less chaotic, but still, important information remains: σ^2 is smaller for countries with high entropy, so in that region (if we have to trace a line, we would say for $H_c >= 4.5$) the dynamics is approximately laminar. This distinction among two different dynamical region was pointed out in [35] for the Fitness method, allowing to separate a *Predictable* form an *Unpredictable* one. Remarkably, also with our measure we observe this separation. This region is even more evident if we use RCA approaches to filter out products not revealing comparative advantage in a country's export basket (Figure 2.11).
In particular, the binarized entropy can picture the best predictability region, which is a consequence of the binarized nature of the bi-adjacency matrix and the consequent loss of information. Countries with a high diversification tend to be stable in the revealed comparative advantage of their products, yielding a more stable complexity than the full information entropy. The filtered RCA entropy seems to not have an advantage over the full information entropy.
A density plot of the collinearity measure, with the same bounds for each colorbar, is helpful to understand better the dynamics (Figure 2.6). We obtain similar collinearity measures for the full information entropy and the filtered RCA one, while the binarized RCA entropy shows a less chaotic behavior.

Entropy construction allows coarse-graining products to get, in principle, more stable country's complexity measures. We observe the dynamics of the coarse-grained entropy (Figure 2.13) by performing the same analysis using the 4 digits aggregation on products (the first level of aggregation offered by the Harmonized System). In the following we will try to understand if we get a less chaotic entropy measure using an aggregation in product from 6 digits to 4 digits.

Predictability and unpredictability regions Qualitatively, we can set a vertical line to discriminate between two regions with low collinearity and high collinearity, in the same fashion of [35]. For 6 digits full information entropy and filtered RCA one we take a line at $H_c = 4.5$ while for binarized RCA entropy we consider $H_c = 6$; instead, for 4 digits full information entropy and filtered RCA one we take a line at $H_c = 4$ while for binarized RCA entropy we consider $H_c = 5$. Computing the average σ^2 in these two different regions will return a

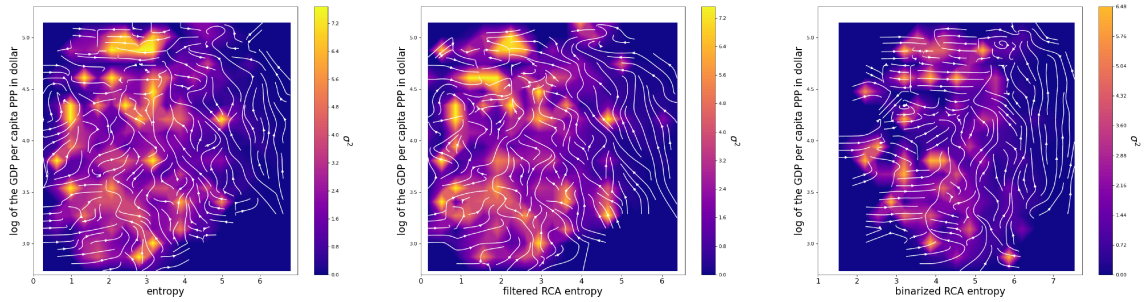


Figure. 2.6: Countour density plot of σ^2 for the different measures. The higher bound of each colorbar is the same. A streamplot is visible to capture better the underlying dynamics.

measures	Unpredictable σ^2	Predictable σ^2	Total σ^2
entropy	3.3 ± 0.3	0.8 ± 0.2	2.5 ± 0.2
filtered RCA entropy	2.8 ± 0.2	1.0 ± 0.3	2.4 ± 0.2
binarized RCA entropy	2.5 ± 0.3	0.5 ± 0.2	1.8 ± 0.2
4 digits entropy	3.2 ± 0.3	0.8 ± 0.1	2.5 ± 0.2
4 digits filtered RCA entropy	3.4 ± 0.4	0.9 ± 0.2	2.8 ± 0.3
4 digits binarized RCA entropy	2.9 ± 0.3	0.57 ± 0.09	2.0 ± 0.2

Table. 2.1: For 6 sigits entropy the predictable region is defined for $H_c > 4.5$ ($H_c > 6$ for binarized RCA entropy), instead for 4 digits entropy the definitions are $H_c > 4$ ($H_c > 5$ for binarized RCA entropy). A distinction in collinearity is present.

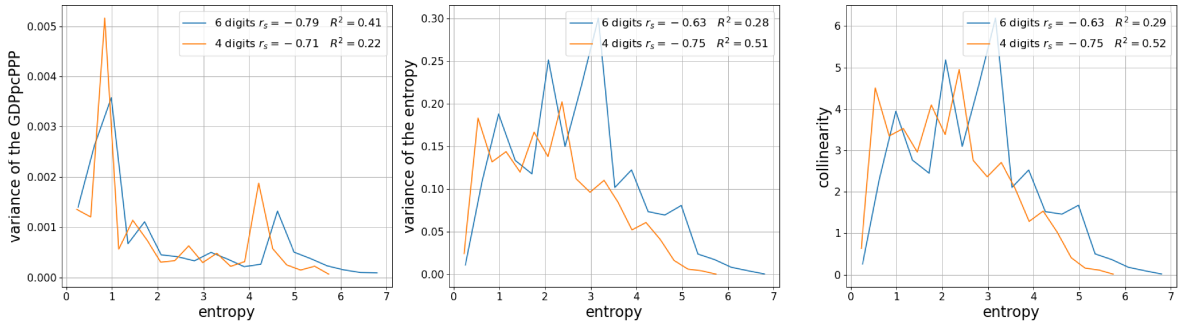


Figure. 2.7: Variance of GDPpcPPP growth, variance of entropy, and collinearity. On the x-axis we find the full information entropy. r_S is the Spearman coefficient, while R^2 is computed using a linear fit.

quantitative insight about the level of chaos in these two regions. We collected this information in a table 2.1.

There is a clear separation in collinearity among these two regions, as we get three times higher average σ^2 in the unpredictability region with respect to the predictable one. Remarkably, the binarized RCA entropy shows lower collinearity in both regions. If we analyze the role of macro aggregation in products for collinearity, we observe that for the 4 digits RCA entropies we get higher σ^2, especially in the unpredictable region, while the full information entropy is not affected by this change. Full information entropy is more stable than RCA entropies when we aggregate products, and this corroborates its use against the other measures with RCA.

Variability of the dynamics as a function of entropy To take the information about entropy and GDPpcPPP variations as a function of entropy, we collect all the coarse-grained boxes that share the same middle entropy and compute the mean variance of their trends. In this way, we can smooth out the noise focusing only on the average contribution of entropy to those quantities. We do the same thing for collinearity.
From Figure 2.7 we observe that low entropy countries have a high variance in GDPpcPPP; the same thing happens in a middle region of entropy, in between 4 and 5, indicating a relatively higher variance in growth for middle-developed countries at the border of the predictability region ($H_c = 4.5$ for 6 digits entropy). The variance in entropy is much more significant than the one we observe in GDPpcPPP, indicating that the collinearity is entirely overwhelmed by entropy dynamics. We find a negative correlation in both the axes, corroborating the idea that high entropic countries have a more stable economy. We studied also the variance of the two axis for the other entropy measures finding that the negative correlation still remains (figure 2.8). Remarkably, using the binarized entropy we do not find the local peak in the middle developed region of countries.

The 4 digits entropy has a different range, as it is bounded from above ($\log(N)$) by a smaller amount of products, this explain the translation of the peaks in the graphs. The similar results in σ^2 in table 2.1, and the comparable shape of the graphs in Figure 2.7 and Figure 2.8, suggest that with the aggregation in products from 6 to 4 digits we do not obtain a less chaotic dynamics.

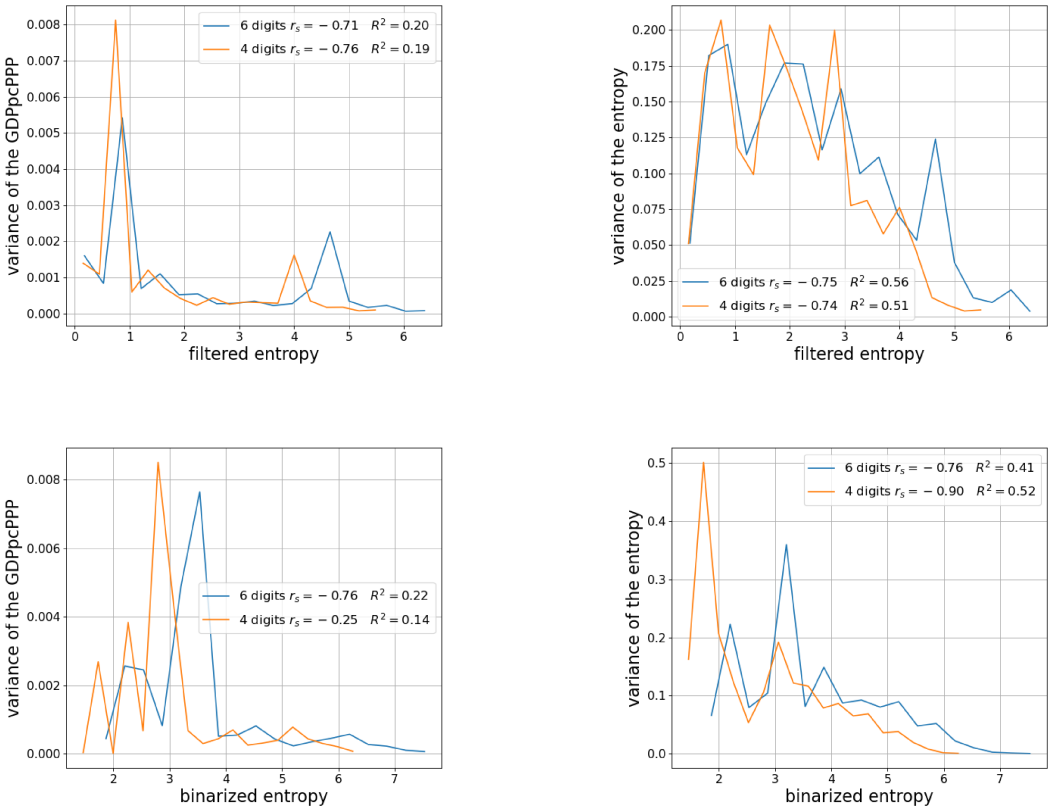


Figure. 2.8: Variance of GDPpcPPP growth, variance of entropy. On the x-axis we find respectively the filtered RCA entropy and the binarized RCA entropy. r_S is the Spearman coefficient, while R^2 is computed using a linear fit.

2.1.1 1, 5 and 10 years trend

To further study the dynamics of the flow, we can augment the temporal range on which we compute velocities. For economic purpose, it is interesting to study middle and long term dynamics; in this sense we repeat the analysis for 5 Figure 2.14 and 10 years Figure 2.15, by computing the velocities using (2.9) and (2.10). We previously observed that we do not obtain a less chaotic plane using the 4 digits entropy, so for this analysis we will only consider the 6 digits one.

$$\{\mathbf{x}_{c,t}\}_b \longrightarrow \{\mathbf{v}_{c,t}\}_b = \{\mathbf{x}_{c,t+5} - \mathbf{x}_{c,t}\} \quad \mathbf{x}_{c,t} \text{ into the } b \text{ coarse-grained box} \tag{2.9}$$

$$\{\mathbf{x}_{c,t}\}_b \longrightarrow \{\mathbf{v}_{c,t}\}_b = \{\mathbf{x}_{c,t+10} - \mathbf{x}_{c,t}\} \quad \mathbf{x}_{c,t} \text{ into the } b \text{ coarse-grained box} \tag{2.10}$$

5 Year Trend

measures	Unpredictable σ^2	Predictable σ^2	Total σ^2
entropy	7.1 ± 0.7	2.9 ± 0.3	5.4 ± 0.5
filtered RCA entropy	6.8 ± 0.6	2.8 ± 0.4	5.4 ± 0.4
binarized RCA entropy	4.5 ± 0.4	1.1 ± 0.3	3.3 ± 0.3

Table. 2.2: Different value of σ^2 for different entropies, highlighting the two regions of predictability

10 Year Trend

measures	Unpredictable σ^2	Predictable σ^2	Total σ^2
entropy	7.9 ± 0.7	4.1 ± 0.5	6.3 ± 0.5
filtered RCA entropy	8.8 ± 0.9	4.2 ± 0.6	7.2 ± 0.6
binarized RCA entropy	5.6 ± 0.6	1.6 ± 0.3	4.1 ± 0.4

Table. 2.3: Different value of σ^2 for different entropies, highlighting the two regions of predictability

As expected, the tables show an increase in σ^2, indicating more chaotic dynamics for middle and long term views. In addition to the analysis made in the previous section, we consider also the logarithmic growth in GDPpcPPP as a function of entropy. As growth can be very heterogeneous for countries sharing similar entropy but different GDPpcPPP, there is the possibility to get outliers in their distribution. Therefore, we also computed the median of the median growth (see equation (2.6)), to smooth out the effect of these outliers. For the 5 year growth Figure 2.9 we obtain an encouraging positive correlation that gets stronger if we look at the long term growth. The correlation at 10-year growth Figure 2.10 between entropy and growth is very high (Spearman coefficient $r_S = 0.82$) and an $R^2 = 0.66$ suggests that a linear relation can in average explain the 66% of the economic growth of a nation. Surprisingly, we get a better correlation and linear fit if we look at the median.
A comparison with the binarized RCA entropy measures on 10-year growth unveils a better correlation using the full information entropy Figure 2.11. Countries with very low binarized entropy are not described well by the measure, as they move away significantly from the 10 years growth trend, corroborating our hypothesis that a diversification in revealed comparative advantage is not a good indicator for under-developed countries. The same consideration for the variance in GDPpcPPP and entropy remain for the middle and long term view, as we can see from the graphs.
It is interesting to notice that we obtain very similar dynamics using the filtered RCA entropy.

We can conclude that the RCA filters the data that contribute most to the entropy measure. Still, at the same time, the full information entropy is stable to the noisy data, capturing essentially the same information as the RCA.

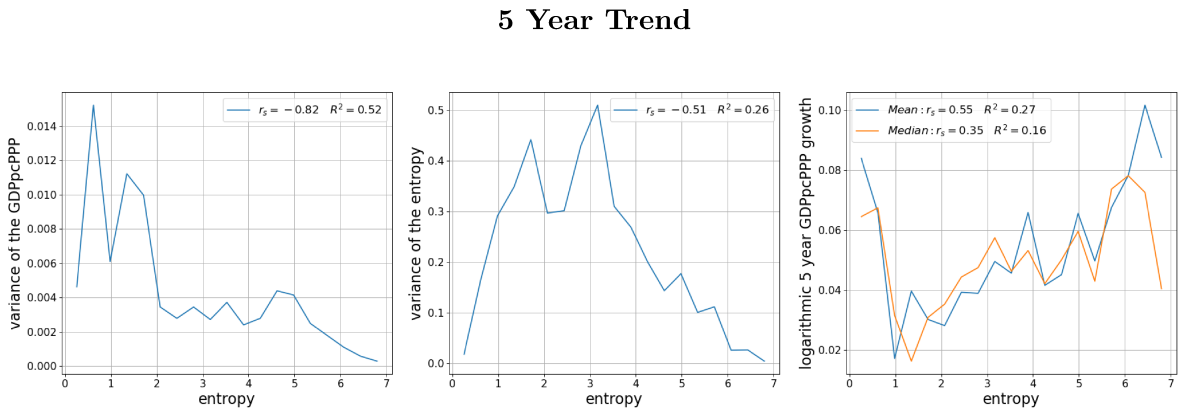


Figure. 2.9: 5 year variance in GDPpcPPP, entropy and logarithmic growth

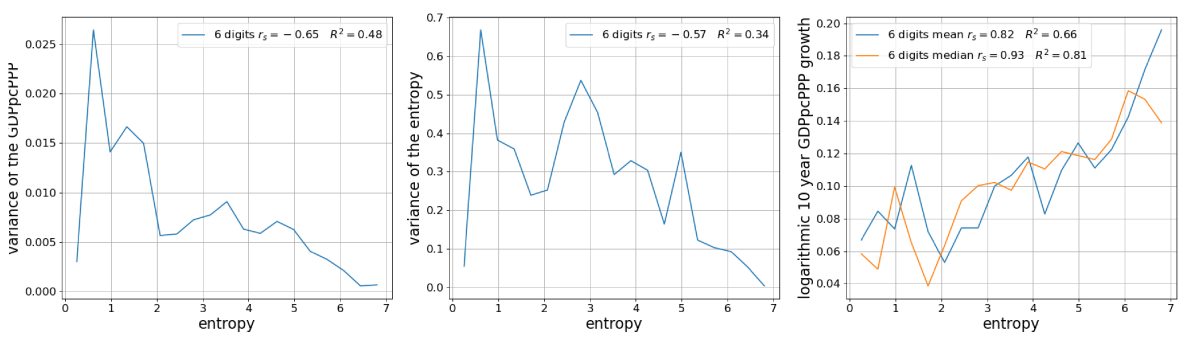


Figure. 2.10: 10 year variance in GDPpcPPP, entropy and logarithmic growth

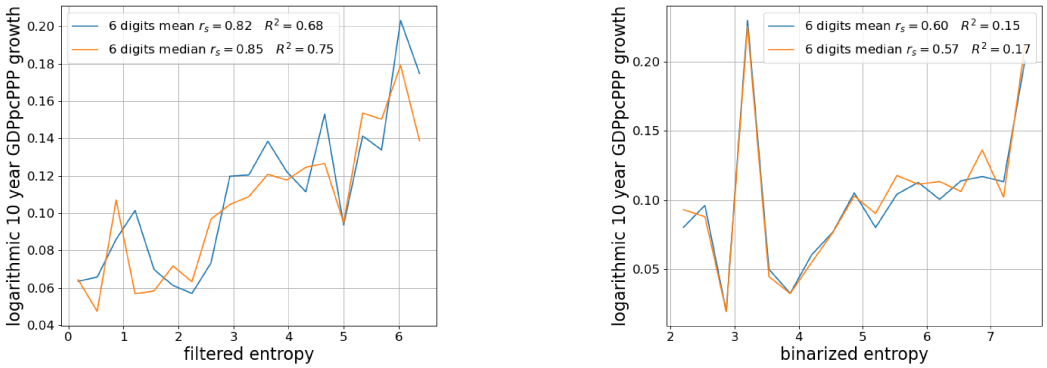


Figure. 2.11: 10 year logarithmic growth for filtered RCA entropy and binarized RCA entropy.

2.2 Patterns in the Macroeconomic Landscape

We better highlight the different dynamical regions that can be found for the 5-year trend, by removing countries with a population smaller than 1 million. Together with the vertical predictability line at $H_c = 4.5$ we plot also the linear trend of entropy-GDPpcPPP.
Interestingly, in the region of risky economy, we get smaller σ^2 if we remove small countries (this can be seen by confronting Figure 2.14 with Figure 2.12), indicating that the most chaos found in that region is due to them. We can argue that small countries with high GDPpcPPP have a high entropy variance because their population constrains diversification. For instance, a nation like San Marino or Monaco cannot reach the level of diversification of countries like Italy or Germany, simply because they do not have the labor capacity to produce and export a high number of products. Moreover, their small population yields a high relative export variability, causing high variance in entropy. In this sense, this type of country is fated to remain in high-income and low entropy regions.
We find an high stability for well developed countries, while we observe a considerable growth for emergent countries. In contrast, under developed countries are affected by an high unpredictability of their economy.

In the next chapter, we will investigate a new algorithm to infer GDPpcPPP growth using the heterogeneous dynamics found in this chapter. We will only use the full information entropy at 6 digits all the dataset regardless country's population.

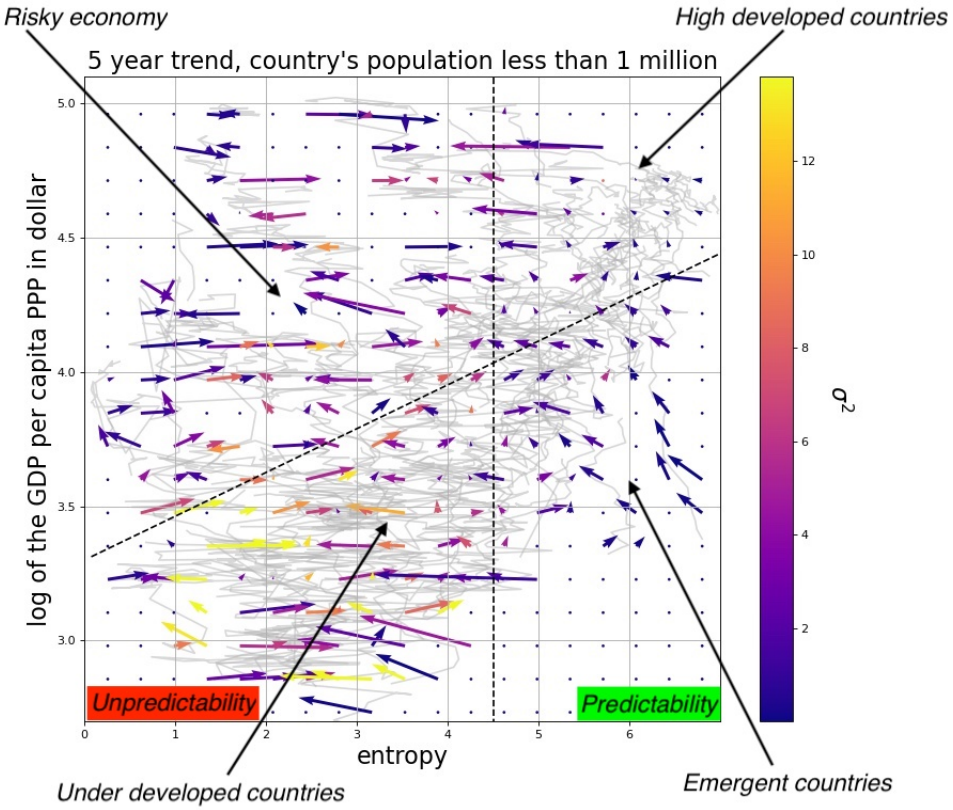


Figure. 2.12: Flows in the entropy-export per capita plane

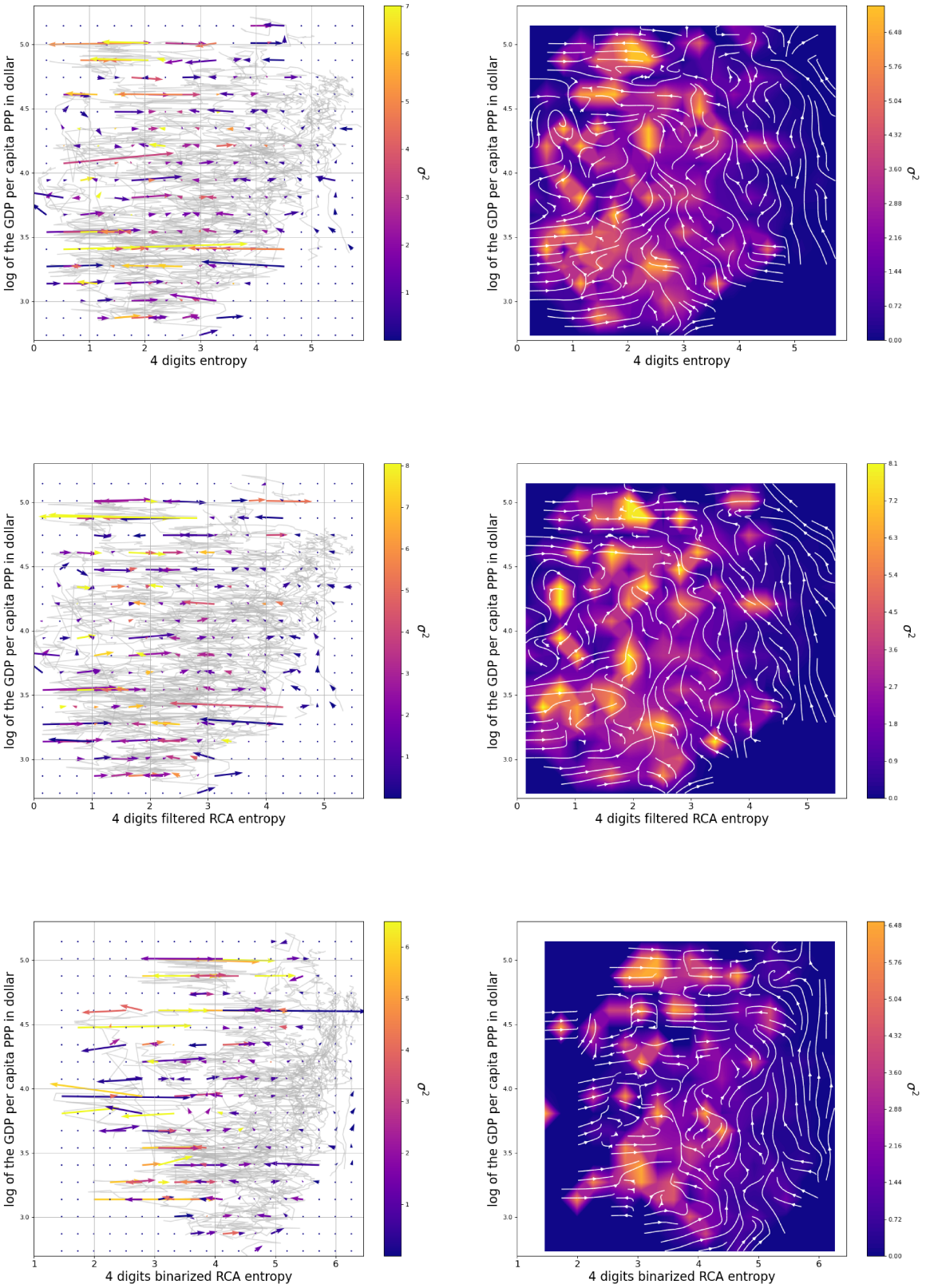


Figure. 2.13: Flow in the entropy - GDPpcPPP plane, with **1 year trend**, using 4 digits aggregation in products. Respectively we observe full information entropy, filtered RCA entropy and binarized RCA entropy.

5 Year Trend

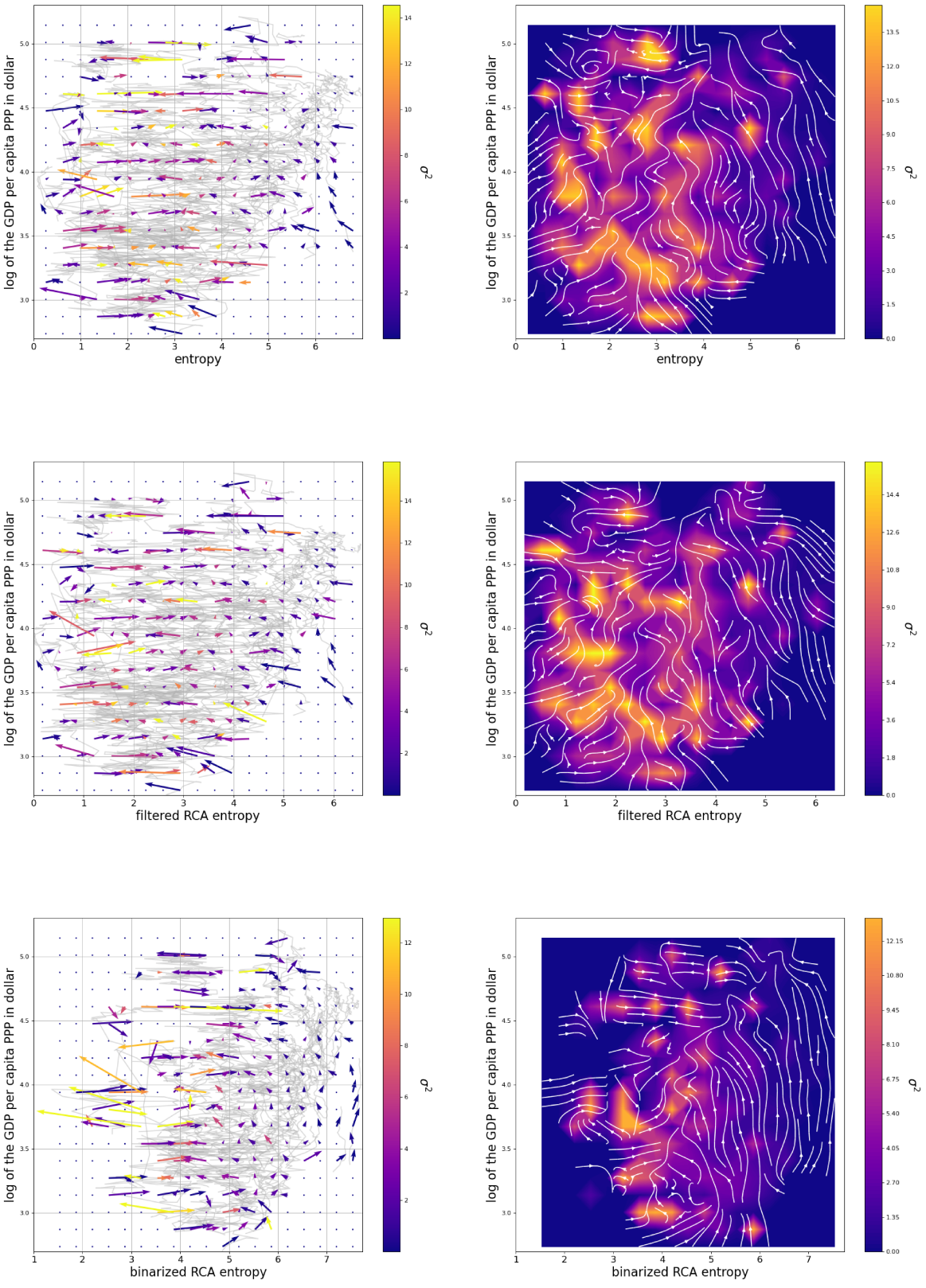


Figure. 2.14: Flow in the entropy - GDPpcPPP plane, with **5 year trend** using 6 digits aggregation in products. Respectively we observe full information entropy, filtered RCA entropy and binarized RCA entropy.

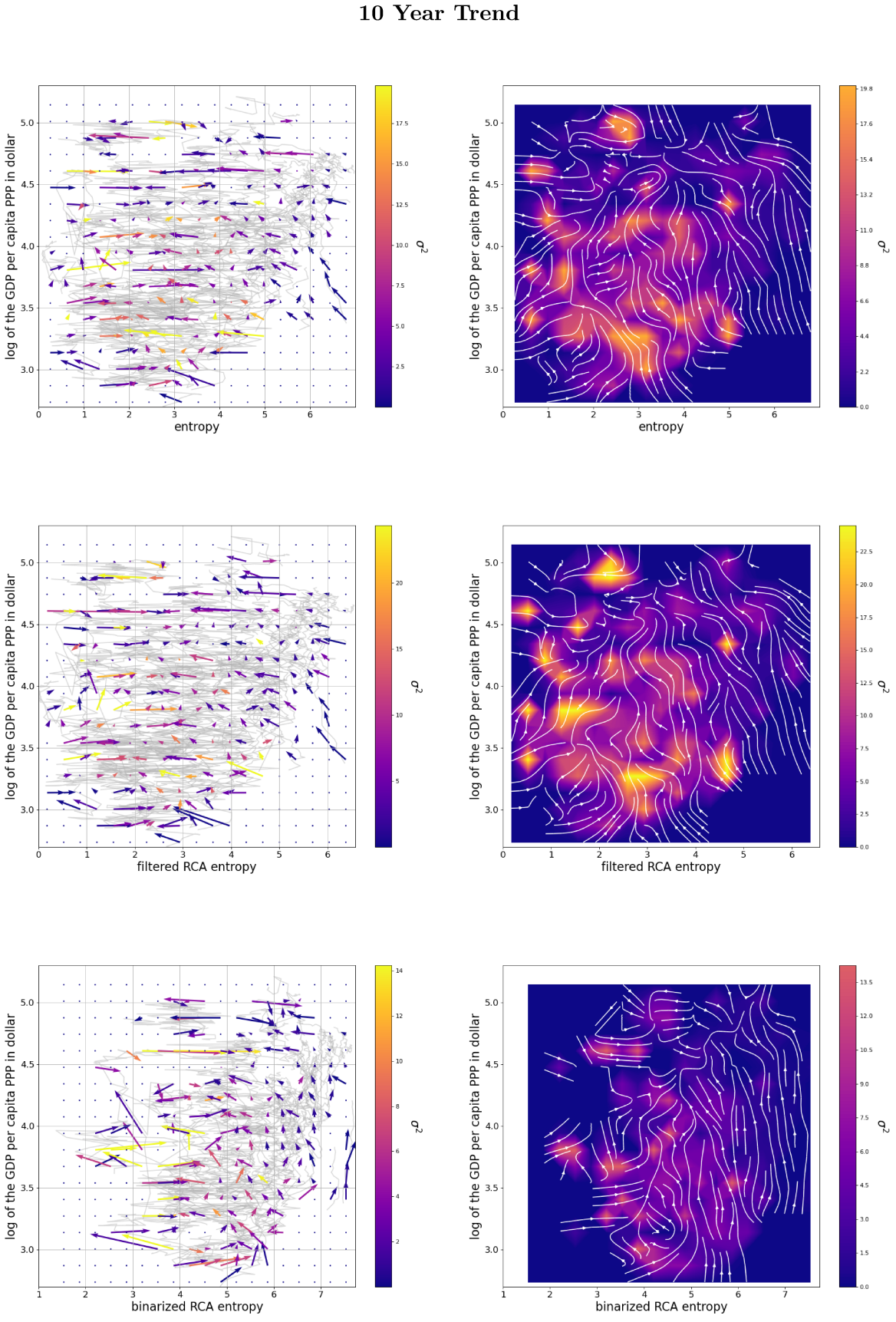


Figure. 2.15: Flow in the entropy - GDPpcPPP plane, with **10 year trend**, using 6 digits aggregation in products. Respectively we observe full information entropy, filtered RCA entropy and binarized RCA entropy.

GDP forecasting

Middle and long-term GDP predictions influence many aspects of a country: a politician might use GDP forecasts to understand how policies will impact the economy, investors diversify their portfolio according to the future economic status of countries. The international monetary fund (IMF) releases GDP projections every year, using complex models that rely on many variables ranging from socio-economics indicators to financial ones. The functional form of the forecast and the parameters on which is based are designed to deliver the best possible prediction. So one ends up with a set of parameters that offer the illusion to catch all the components that drive economic growth, which is actually very heterogeneous. For example, it is scarcely interpretable a possible relation between population age and raw material prices. Within the economic complexity framework, we aim to avoid this abundance of parameters by concentrating only on the complexity measure of nations and their GDP.
In this context, we developed a new complexity measure and showed, in the second chapter, that a coupling with GDPpcPPP identifies dynamical patterns in the macro-economic landscape. In particular, we found that countries close in the plane evolve similarly in the predictability region. In principle, we can infer a country's future dynamics by looking at what was historically observed in its neighborhood on the entropy-GDP plane. More specifically, given a point $\mathbf{x}$ in the plane that describe a nation in a particular year, looking at the evolutions of points closed to $\mathbf{x}$ will give some insights about the evolution of the point $\mathbf{x}$. The points close to $\mathbf{x}$ with known evolutions are called *analogues*, and the approach goes under the name of *Method of Analogues* (developed by E.N. Lorents in the context of atmospheric predictability [37]). The laminar dynamics observed in the predictability region of the entropy-GDPpcPPP plane justifies this method for forecasting GDP growth, therefore using a less complex and more interpretable model than the ones used by the IMF. The method's accuracy depends only on the recipe we use to choose analogues and the procedure for extracting the prediction.
In this chapter we will investigate and reformulate the *Selective Predictability Scheme bootstrap*, a recipe of the method of analogues developed for the fitness complexity measure [36], applying it to the entropy-GDPpcPPP plane. To avoid inflation we aim to predict growth in GDPpcPPP instead of GDP, therefore to simplify the notation from now on we will write GDP instead of GDPpcPPP.

3.1 Selective Predictability Scheme bootstrap

Initially proposed in [36], the idea that countries with similar GDP and Fitness evolve similarly was exploited using the *Method of Analogue*. The log fitness - log GDP plane was embedded with a Euclidean distance to measure the closeness of analogues. In addition, to extract the information from them they decided to sample analogues from a univariate Gaussian probability distribution and then calculate the mean growth based on the set of samples.
We can summarize the selective predictability scheme bootstrap (SPSb) in the following steps.

1. Given an initial point $\mathbf{x}_{c,t}$ in the log fitness - log GDP plane (where c indicates a country and t a specific year) we want to forecast its 5 year displacement vector, for $\Delta t = 5$.

$$\delta\mathbf{x}_{c,t} = \mathbf{x}_{c,t+\Delta t} - \mathbf{x}_{c,t} \tag{3.1}$$

The analogues, which are points "close" to $\mathbf{x}_{c,t}$, are indicated as $\mathbf{x}_{\tilde{c},\tau}$ with $\tau + \Delta t \leq t$ to guarantee that we are not using analogue's evolution $\mathbf{x}_{\tilde{c},\tau+\Delta t}$ that would not be known at time t. The pedix $\tilde{c}$ indicates that the point $\mathbf{x}$ can be associated with any country in the dataset.

2. We sample with repetition N analogues (where N is the number of available analogues) according to a univariate Gaussian conditional probability density function, getting a set $\{\mathbf{x}^i_{\tilde{c},\tau}\}_{i=1,\dots N}$

$$p(\mathbf{x}_{\tilde{c},\tau}|\mathbf{x}_{c,t}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{|\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}|^2}{2\sigma^2} \right) = p(\ell) \tag{3.2}$$

Where σ^2 is the variance of the probability distribution and $\ell = |\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}|$ the Euclidean distance. The sampling probabilities $P(\mathbf{x}_{\tilde{c},\tau}|\mathbf{x}_{c,t})$ are then given by the cumulative of this function.

$$P(\mathbf{x}_{\tilde{c},\tau}|\mathbf{x}_{c,t}) = P(\ell) = 1 - \int_0^\ell p(\tilde{\ell})\mathrm{d}\tilde{\ell} \tag{3.3}$$

A possible sampling technique could be a simple accept-reject algorithm [1]. We decide to consider a point $\mathbf{x}_{\tilde{c},\tau}$ an analogue of $\mathbf{x}_{c,t}$ if its probability to be sampled is larger than 0.05.

3. For each sampled analogue we take its 5 year displacement vector

$$\mathbf{x}^i_{\tilde{c},\tau} \mapsto \delta\mathbf{x}^i_{\tilde{c},\tau} = \mathbf{x}^i_{\tilde{c},\tau+\Delta t} - \mathbf{x}^i_{\tilde{c},\tau} \tag{3.4}$$

and compute the average displacement.

$$\langle\delta\mathbf{x}_{c,t}\rangle_B = \frac{1}{N}\sum_{i=1}^{N} \delta\mathbf{x}^i_{\tilde{c},\tau} \tag{3.5}$$

4. We repeat the second and third step for $M = 1000$ times (this is the bootstrap procedure), obtaining a distribution of M different displacement $\{\langle\delta\mathbf{x}_{c,t}\rangle_B\}_{B=1\dots M}$.

5. We take as final prediction the mean of the distribution given by the bootstrap and as error its variance.

$$\langle\delta\mathbf{x}_{c,t}\rangle = \frac{1}{M}\sum_{B=1}^{M}\langle\delta\mathbf{x}_{c,t}\rangle_B \qquad \sigma^2_{\langle\delta\mathbf{x}_{c,t}\rangle} = \frac{1}{M}\sum_{B=1}^{M}(\langle\delta\hat{\mathbf{x}}_{\mathbf{c},\mathbf{t}}\rangle_B - \langle\delta\hat{\mathbf{x}}_{\mathbf{c},\mathbf{t}}\rangle)^2 \tag{3.6}$$

6. The forecast is then

$$\langle\mathbf{x}_{c,t+\Delta t}\rangle = \mathbf{x}_{c,t} + \langle\delta\mathbf{x}_{c,t}\rangle \tag{3.7}$$

This method has shown significant improvement for GDP forecasts as it outperforms the predictions made by the IMF. In this section, we will deeply investigate this algorithm focusing on some improvement, but more importantly, it will be applied in the entropy-GDP plane showing that with our measure we can outperform IMF predictions as well.

[1] In our simulations we considered this sample algorithm: we randomly pick an analogue with uniform distribution, we compute its probability $P(\ell)$ according to (3.3), we generate a random number p between 0 and 1 and we accept the analogue if $P(\ell) \geq p$.

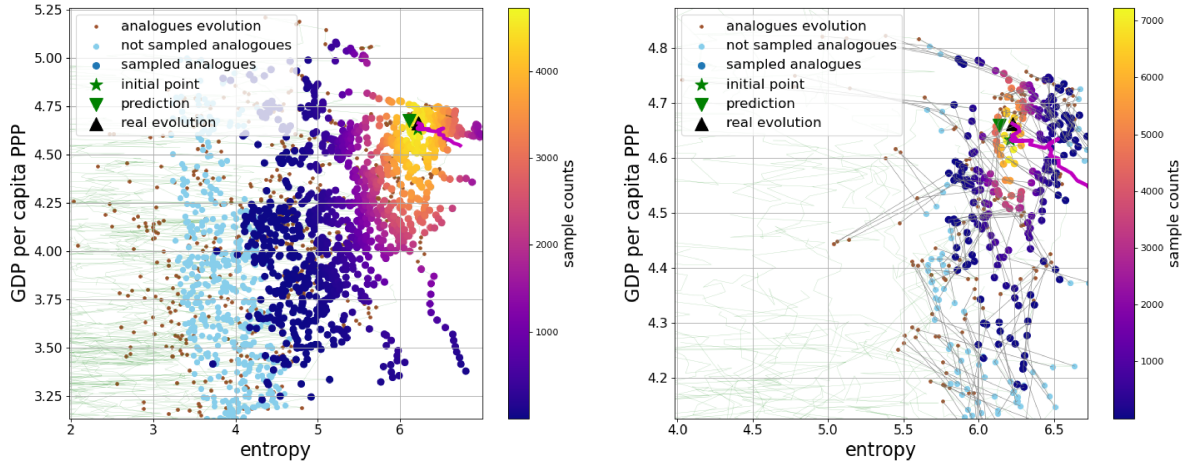


Figure. 3.1: SPSb for France, prediction from 2014 to 2019. In the left plot we have $\sigma = 0.5$, while in the right one $\sigma = 0.1$. In the right graph we have a zoomed situation around the initial point, to highlight the displacements of the analogues as gray lines. Small brown dots represent the evolution and the sampled analogues are depicted using a colorbar indicating their relevance, while cyan dots are analogues that have not been sampled. The purple line indicates the France time series.

3.1.1 Observations and critiques of the algorithm.

At first sight, it is immediate to observe a problem; how do we define the variance σ^2 in (3.2)? With a high variance, we would sample many analogues with no criterion, while with a small variance, we could end up with a very small data-set of analogue's evolution, possibly causing a biased estimate. For example, we can consider the prediction of GDP growth of France from 2014 to 2019 (figures 3.1): in the left graph, we used a standard deviation $\sigma = 0.5$ while in the right one we considered a smaller one $\sigma = 0.1$. The predictions are different in these two cases. In the original paper [36], a standard deviation of $\sigma = 0.5$ was proposed with no explanation, however a criterion to set the variance is required.

More parameters Another important observation is the following, why do we have to take a Euclidean distance in the plane? There is no way to decide a priori that a circle around a point in the economic plane contains analogues sharing the same importance, so the same distance. In this sense, we decided to improve the probability computation by choosing a multivariate Gaussian probability distribution.

$$p(\mathbf{x}_{\tilde{c},\tau}|\mathbf{x}_{c,t}) = \frac{1}{2\pi\sqrt{|\det(\mathbf{H})|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}\right)^T \mathbf{H}^{-1}\left(\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}\right)\right] \tag{3.8}$$

With a general covariance matrix $\mathbf{H}$ we have three parameters to set.

$$\mathbf{H} = \begin{pmatrix} \mathrm{Var}[x] & \mathrm{Cov}[x,y] \\ \mathrm{Cov}[x,y] & \mathrm{Var}[y] \end{pmatrix} \qquad \text{with } (x,y) = \mathbf{x} \tag{3.9}$$

In the case of zero covariance, the interpretation would be simple. A covariance matrix with $\mathrm{Var}[x] > \mathrm{Var}[y]$ will consider countries with similar GPCpcPPP closer than countries with similar entropy; it is true also the opposite.
With the multivariate Gaussian approach, we can define a generalized distance in the economic

plane, the Mahalanobis distance [53].

$$r(\mathbf{x}_{\tilde{c},\tau}, \mathbf{x}_{c,t}) = \sqrt{\left(\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}\right)^T \mathbf{H}^{-1} \left(\mathbf{x}_{\tilde{c},\tau} - \mathbf{x}_{c,t}\right)} \tag{3.10}$$

Points having the same Mahalanobis distance lay in an ellipse of equal conditional probability. In order to assign probabilities to the analogues we have to define a cumulative probability function of the multivariate Gaussian distribution. Since analogues with the same Mahalanobis distance from the initial point are considered equally distant, we define the cumulative probability function of the multivariate Gaussian distribution as the area of the ellipse (locus of points with equal Mahalanobis distance)[53].

$$F(r) = 1 - e^{-r^2/2} \tag{3.11}$$

Therefore, the probability associated with each analogue is complementary to this cumulative function, as ellipses with small areas contain points closer to the initial point.

$$P(\mathbf{x}_{\tilde{c},\tau} | \mathbf{x}_{c,t}) = e^{-r(\mathbf{x}_{\tilde{c},\tau}, \mathbf{x}_{c,t})^2/2} \tag{3.12}$$

In this context the SPSb works in the same way as it was introduced, with the only difference on how we compute probabilities.
To get the original SPSb in this multivariate formulation is sufficient to consider a diagonal covariance matrix with equal elements, hence $\mathrm{Var}[x] = \mathrm{Var}[y] = \sigma^2$ and $\mathrm{Cov}[x, y] = 0$. We will analyze in the next sections which is the more appropriate model, referring to multivariate SPSb and univariate SPSb.

The next step is to find a criterion to set the parameters of the covariance matrix $\mathbf{H}$. SPSb is intrinsically slow, as it relies on sampling with repetition, making an optimization procedure intractable for a personal computer, especially if we use a multivariate Gaussian probability. In the next section, we will see how SPSb converges to a faster algorithm.

3.2 Convergence of the Algorithm

The SPSb algorithm is sensitive to the choice of the sampling number N and the number of bootstrap M, but more importantly, it requires an extensive computational effort making the optimization approach intractable. The mean displacement $\langle \delta \mathbf{x}_{c,t} \rangle$ is independent of the bootstrap's procedure, but it only depends on how many analogues we sample.

$$\langle \delta \mathbf{x}_{c,t} \rangle = \frac{1}{M} \sum_{B=1}^{M} \langle \delta \mathbf{x}_{c,t} \rangle_B = \frac{1}{M \cdot N} \sum_{i=1}^{N \cdot M} \delta \mathbf{x}^i_{\tilde{c},\tau} \tag{3.13}$$

Where the last sum is intended over the different analogues of $\mathbf{x}_{c,t}$ labeled as $(\tilde{c}, \tau)$ among the M batches made of N samples. To simplify the notation from now on we are going to use a multi-index to identify analogues $\mathcal{C} = (\tilde{c}, \tau)$.
Sticking to the example of France's growth from 2014 to 2019, to investigate the algorithm's convergence for a large number of samples we made 20 different simulations using the same standard deviation $\sigma = 0.1$, as shown in Figure 3.2. The algorithm seems to converge to a unique value. Calling the total number of samples $\mathcal{N}$ (so in (3.13) we have $\mathcal{N} = N \cdot M$) and rewriting the average of the analogue's displacement in the limit of infinite bootstrap as a weighted average over the absolute sample's frequency $n_{\mathcal{C}}$ we get

$$\langle \delta \mathbf{x}_{c,t} \rangle = \lim_{\mathcal{N} \to \infty} \frac{\sum_{\mathcal{C}}^{N} n_{\mathcal{C}} \delta \mathbf{x}_{\mathcal{C}}}{\mathcal{N}} = \lim_{\mathcal{N} \to \infty} \sum_{\mathcal{C}}^{N} \left(\frac{n_{\mathcal{C}}(\mathcal{N})}{\mathcal{N}} \right) \delta \mathbf{x}_{\mathcal{C}} \tag{3.14}$$

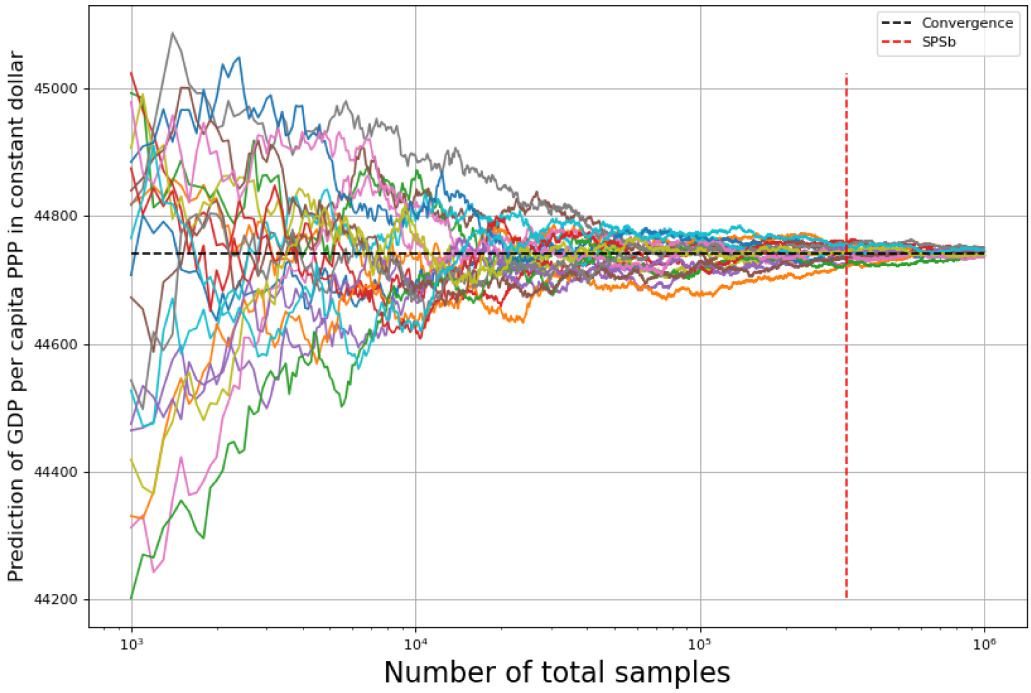


Figure. 3.2: Convergence of SPSb, the red line indicate the SPSb stop.

Where the sum is intended over the N different analogues of $\mathbf{x}_{c,t}$ and in the last term, we explicitly write the dependence of $n_{\mathfrak{C}}$ from $\mathcal{N}$. In a frequentist view the last coefficient should converge to a vector of probabilities.

$$\lim_{\mathcal{N}\to\infty}\left(\frac{n_{\mathfrak{C}}(\mathcal{N})}{\mathcal{N}}\right) = p_{\mathfrak{C}} \tag{3.15}$$

Where $p_{\mathfrak{C}}$ is the probability to sample the analogue $\mathbf{x}_{\mathfrak{C}}$ with respect to the other N analogues. This probability should be related to the cumulative distribution function we have defined for the Mahalanobis distance. The difference is that the function (3.12) gives the probability of sampling an analogue laying in an ellipse, while we are interested in knowing the sampling probability marginalized to a restricted set of N analogues. So intuitively, this probability should be given by the cumulative function normalized with respect to the total probability.

$$P_{\mathfrak{C}} = e^{-r_{\mathfrak{C}}^2/2} \quad \text{Probability to sample i-th analogue laying in an ellipse of Mahalanobis distance } r_i$$

$$p_{\mathfrak{C}} = \frac{P_{\mathfrak{C}}}{\sum_{\mathfrak{C}}^{N} P_{\mathfrak{C}}} \quad \text{Probability to sample i-th analogue with respect to the other N analogues}$$

The equality is confirmed through to simulations shown in Figure 3.2.

$$\lim_{\mathcal{N}\to\infty}\left(\frac{n_{\mathfrak{C}}(\mathcal{N})}{\mathcal{N}}\right) = \frac{P_{\mathfrak{C}}}{\sum_{\mathfrak{C}}^{N} P_{\mathfrak{C}}} \tag{3.16}$$

Therefore, we decided to replace to SPSb algorithm with the convergent one (we will call it cSPS). The new estimates are now given by the following formulas

$$\langle \delta \mathbf{x}_{c,t} \rangle_{\text{cSPS}} = \frac{\sum_{\mathcal{C}}^{N} e^{-r(\mathbf{x}_{\mathcal{C}}\,,\,\mathbf{x}_{c,t})^2/2} \delta \mathbf{x}_{\mathcal{C}}}{\sum_{\mathcal{C}}^{N} e^{-r(\mathbf{x}_{\mathcal{C}}\,,\,\mathbf{x}_{c,t})^2/2}} \tag{3.17}$$

This regression is similar to a non-linear regression technique called Nadaraya-Watson kernel regression. The similarity was studied in [34] with the SPSb, but the multivariate case presented here represents an original improvement. More information on Nadaraya-Watson kernel regression can be found in the appendix.

3.2.1 Nadaraya Watson applied to SPSb

In this context the selective predictability scheme has a very strong mathematical foundation. We have a number of observation $\{(\mathbf{x}_{\mathcal{C}}\,,\,\delta\mathbf{x}_{\mathcal{C}})\}_{\mathcal{C}}$ (remember that $\mathcal{C}$ is a multi-index $\mathcal{C} = (\tilde{c}, \tau)$ that identifies an analogue with a specific time) and given the point $\mathbf{x}_{c,t}$ we want to estimate $\delta\mathbf{x}_{c,t}$. Using the Nadaraya-Watson kernel regression, with a multivariate Gaussian kernel centered on $\mathbf{x}_{c,t}$, leads us to the following estimator

$$\langle \delta \mathbf{x}_{c,t} \rangle_{NW} = \hat{E}(\delta \mathbf{x}_{c,t} | \mathbf{x}_{c,t}) = \frac{\sum_{\mathcal{C}}^{N} K_{\mathbf{H}}(\mathbf{x}_{c,t} - \mathbf{x}_{\mathcal{C}}) \delta \mathbf{x}_i}{\sum_{\mathcal{C}}^{N} K_{\mathbf{H}}(\mathbf{x}_{c,t} - \mathbf{x}_{\mathcal{C}})} = \frac{\sum_{\mathcal{C}}^{N} e^{-r(\mathbf{x}_{\mathcal{C}}\,,\,\mathbf{x}_{c,t})^2/2} \delta \mathbf{x}_{\mathcal{C}}}{\sum_{\mathcal{C}}^{N} e^{-r(\mathbf{x}_{\mathcal{C}}\,,\,\mathbf{x}_{c,t})^2/2}} \tag{3.18}$$

Which is indeed the same estimator found intuitively in (3.17). With this approach, we can also compute the errors of our prediction without using the bootstrap technique.

$$\sigma^2_{\langle \delta \mathbf{x}_{c,t} \rangle_{NW}} = \hat{E}\big[\big(\delta \mathbf{x}_{c,t} - \hat{E}(\delta \mathbf{x}_{c,t} | \mathbf{x}_{c,t})\big)^2\big] \tag{3.19}$$

The estimator's accuracy depends on the choice of the bandwidths of the kernel (the matrix $\mathbf{H}$), hence on the choice of the three parameters of the covariance matrix used to sample analogues.

3.2.2 The cSPS algorithm

For the sake of completeness and clarity, we summarize the cSPS algorithm in the following steps:

1. Given an initial point $\mathbf{x}_{c,t}$ in the log fitness - log GDP plane (where c indicates a country and t a specific year) we want to forecast its 5 year displacement vector, for $\Delta t = 5$.

$$\delta \mathbf{x}_{c,t} = \mathbf{x}_{c,t+\Delta t} - \mathbf{x}_{c,t} \tag{3.20}$$

2. Using a multivariate Gaussian kernel with bandwidth $\mathbf{H}$ and centered on $\mathbf{x}_{c,t}$, we collect a set of analogues $\{\mathbf{x}_{\tilde{c},\tau}\}$ by taking all the points in the entropy-GDP plane with

$$P(\mathbf{x}_{\tilde{c},\tau} | \mathbf{x}_{c,t}) = e^{-r(\mathbf{x}_{\tilde{c},\tau}, \mathbf{x}_{c,t})^2/2} > 0.05 \qquad \text{and} \qquad \tau \leq t - \Delta t \tag{3.21}$$

   Where $r(\mathbf{x}_{\tilde{c},\tau}, \mathbf{x}_{c,t})$ is the Mahalanobis distance computed in (3.10), which depends on the bandwidth $\mathbf{H}$. While the first condition is set only for computational convenience, the second condition on the time τ is important to a strict out-of-sample prediction.

3. For each sampled analogue we take its Δt year displacement vector

$$\mathbf{x}_{\tilde{c},\tau} \mapsto \delta \mathbf{x}_{\tilde{c},\tau} = \mathbf{x}_{\tilde{c},\tau+\Delta t} - \mathbf{x}_{\tilde{c},\tau} \tag{3.22}$$

4. We estimate the unknown evolution $\delta\mathbf{x}_{c,t}$ with a Nadaraya-Watson kernel regression technique over the set of the analogues' evolution

$$\langle\delta\mathbf{x}_{c,t}\rangle_{\mathrm{NW}} = \frac{\sum_{(\tilde{c},\tau)}^{N} e^{-r(\mathbf{x}_{\tilde{c},\tau},\,\mathbf{x}_{c,t})^2/2}\delta\mathbf{x}_{\tilde{c},\tau}}{\sum_{(\tilde{c},\tau)}^{N} e^{-r(\mathbf{x}_{\tilde{c},\tau},\,\mathbf{x}_{c,t})^2/2}} \tag{3.23}$$

The error of the estimate is given again by the kernel regression

$$\sigma^2_{\langle\delta\mathbf{x}_{c,t}\rangle_{\mathrm{NW}}} = \frac{\sum_{(\tilde{c},\tau)}^{N} e^{-r(\mathbf{x}_{\tilde{c},\tau},\,\mathbf{x}_{c,t})^2/2}\big(\delta\mathbf{x}_{\tilde{c},\tau} - \langle\delta\mathbf{x}_{c,t}\rangle_{\mathrm{NW}}\big)^2}{\sum_{(\tilde{c},\tau)}^{N} e^{-r(\mathbf{x}_{\tilde{c},\tau},\,\mathbf{x}_{c,t})^2/2}} \tag{3.24}$$

The role of the bandwidth $\mathbf{H}$ has never been discussed in the literature. We will address this problem in the next section.

3.3 Bandwidth Selection

To compute the error in GDP predictions, we choose the difference in compound annual growth rate (CAGR) in the same way as the reference paper [36].

$$\mathrm{CAGR}\% = \left[\left(\frac{\text{final value}}{\text{initial value}}\right)^{1/\Delta t} - 1\right]100\% \tag{3.25}$$

The errors are therefore computed as deviation from the true CAGR.

$$E = \mathrm{CAGR}\% - \mathrm{CAGR}\%_{\text{forecasted}} \tag{3.26}$$

Using this error we can investigate better the role of bandwidth selection in the accuracy of the cSPS algorithm, using the evolutions from 2014 to 2019 as a benchmark. In the context of statistical learning the considered set $\{(\mathbf{x}_{c,2014}, \delta\mathbf{x}_{c,2014})\}_c$ is called *test set*. The accuracy of the cSPS is expressed with the mean absolute error (MAE) on a total of 95 country's growth forecasts, the same that have counterpart predictions made by the IMF. Using the univariate cSPS, in Figure 3.3 we observe a classical U-shape in $\mathrm{MAE}(\sigma)$ which shows a minimum at $\sigma^* = 0.25$. However, a direct optimization procedure on the test set is not possible, because we have to put ourselves in a situation in which data after 2014 are unknown (so we have to take out-of-sample data).
Under the assumption that the optimal bandwidth does not change significantly over the year, we developed a strict out-of-sample technique to infer the best parameters σ^* for the univariate cSPS and $\mathbf{H}^*$ for the multivariate cSPS, on a *training set* with known evolution at the year of the test set.

Bandwidth for the cSPS With the Nayadara-Watson kernel regression, the estimates become deterministic (3.17), so for each bandwidth, a unique error is given.

$$\sigma : (\sigma,\sigma,0) \to E(\sigma,\sigma,0) = E(\sigma) \qquad \text{Univariate cSPS} \tag{3.27}$$

$$\mathbf{H} : (\sigma_x,\sigma_y,\theta) \to E(\sigma_x,\sigma_y,\theta) = E(\mathbf{H}) \qquad \text{Multivariate cSPS} \tag{3.28}$$

The bandwidth $\mathbf{H}$ in (3.17) in the context of kernel regression, is completely described using a rotated diagonal matrix, hence by three parameters.

$$\mathbf{H} = \begin{pmatrix} \cos^2\theta\sigma_x^2 + \sin^2\theta\sigma_y^2 & \cos\theta\sin\theta\sigma_x^2 - \cos\theta\sin\theta\sigma_y^2 \\ \cos\theta\sin\theta\sigma_x^2 - \cos\theta\sin\theta\sigma_y^2 & \sin^2\theta\sigma_x^2 + \cos^2\theta\sigma_y^2 \end{pmatrix} \tag{3.29}$$

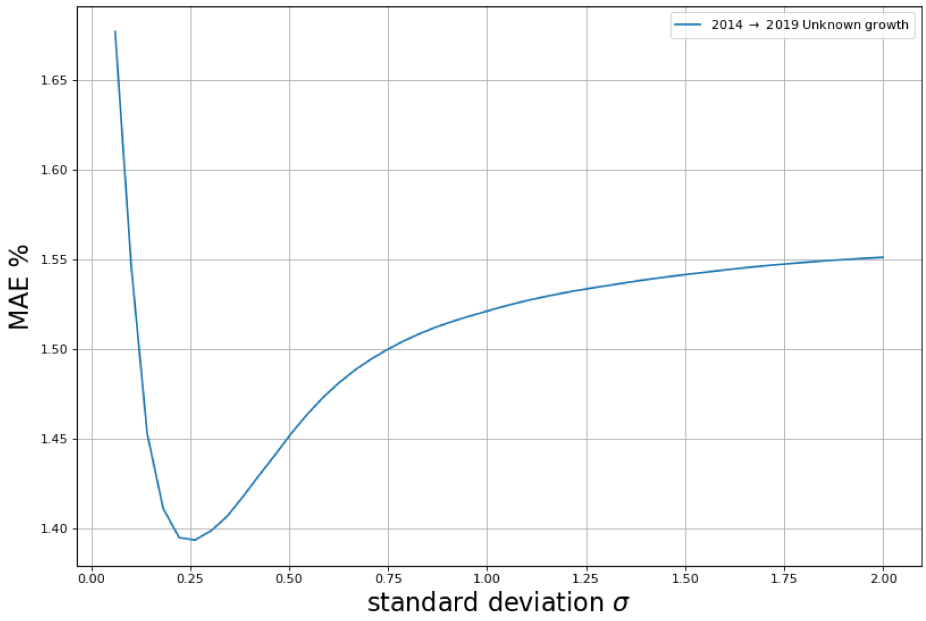


Figure. 3.3: Dependence of MAE to the choice of the single bandwidth σ.

This representation, other than having a clear geometrical interpretation, allows us to safely sample the parameter space $(\sigma_x, \sigma_y, \theta)$ as the transformation to $\mathbf{H}$ always gives a positive definite matrix, which is essential for the multivariate Gaussian kernel. In order to have meaningful parameters according to the method of analogues, we set bounds $\sigma_x \in [0.01, 1]$, $\sigma_y \in [0.01, 1]$ and $\theta \in [-\pi/2, \pi/2]$ in the minimization procedure. For the univariate cSPS we simply sample over a line $\sigma \in [0.01, 1]$.

3.3.1 Statistical Learning for Bandwith Selection

To expose the methods, we will only deal with the multivariate cSPS, since the univariate cSPS is just a particular case.

We developed two different methods to infer the best bandwidth $\mathbf{H}^*$: a unique value that is valid for all the plane (*constant* $\mathbf{H}^*$) and a specialized version, that returns for each country a best bandwidth (*specialized* $\mathbf{H}^*(\mathbf{x}_{c,t})$). To perform the minimization we used the "Limited-memory Broyden–Fletcher–Goldfarb–Shannon with Bounds" (L-BFGS-B) algorithm contained in the scipy library on python. We chose this algorithm as it is the default choice for scipy when we have to deal with bounds in the parameters space.

Constant $\mathbf{H}^*$ We use as training set the most recent known evolutions at the time of the test set.

$$\{(\mathbf{x}_{c,t}, \delta\mathbf{x}_{c,t})\}_c \qquad \text{test set} \tag{3.30}$$
$$\{(\mathbf{x}_{c,t-\Delta t}, \delta\mathbf{x}_{c,t-\Delta t})\}_c \qquad \text{training set} \tag{3.31}$$

We will call t *test time* while $t - \Delta t$ *training time*. Using the training set we learn a unique bandwidth $\mathbf{H}^*$ by minimizing the mean absolute error (MAE). For instance, if we want to

forecast the GDP growth from 2014 to 2019, we can look at previous evolutions from 2009 to 2014 for each country getting a set of errors $\{E_c\}_c$. The minimization is performed on the MAE function

$$\text{MAE}(\mathbf{H}) = \frac{1}{N_c}\sum_c |E_c(\mathbf{H})| \qquad \mathbf{H}^* = \arg\min(\text{MAE}(\mathbf{H})) \tag{3.32}$$

Where N_c is the number of countries considered. The predictions on the training set rely on a slightly reformulation of the cSPS presented in section 3.2.2, especially in the time condition on the second step. Since we want to use the maximum information of the database, hence all the possible analogues with known evolution at the test time, we can reformulate the second step as follow:

> Given the test time t and the training time $t - \Delta t$. Using a multivariate Gaussian kernel with bandwidth $\mathbf{H}$ and centered on $\mathbf{x}_{c,t-\Delta t}$, we collect a set of analogues $\{\mathbf{x}_{\tilde{c},\tau}\}$ by taking all the points in the entropy-GDP plane with
>
> $$P(\mathbf{x}_{\tilde{c},\tau}|\mathbf{x}_{c,t}) = e^{-r(\mathbf{x}_{\tilde{c},\tau},\mathbf{x}_{c,t})^2/2} > 0.05 \qquad \text{and} \qquad \tau \leq t - \Delta t \tag{3.33}$$

With this reformulation, we allow sampling analogues with an evolution not known at the training time but at the test time. For instance, in the training set of 2014 with $\Delta t = 5$ we can now sample analogues with a year less or equal to 2009 (instead of 2004), so the important aspect is that their evolution is known at the test time.
Moreover, with this reformulation the algorithm uses the same information on the training and the test set, hence the same space of possible analogues.

Specialized $\mathbf{H}^*(\mathbf{x}_{c,t})$ A second, more complex way to get the best bandwidth $\mathbf{H}^*$ is to specialize it for each country in the test set. The idea is that the best bandwidth for each country $\mathbf{H}^*(\mathbf{x}_{c,t})$ can be inferred by optimizing it on a training set composed by close analgoues to $\mathbf{x}_{c,t}$. In figure 3.4 are depicted the ten most close analogues to France in the year 2014, the closeness is computed with a Euclidean distance.
We can take K close analogues to $\mathbf{x}_{c,t}$ and perform the optimization on those

$$\text{MAE}(\mathbf{H}(\mathbf{x}_{c,t})) = \frac{1}{K}\sum_{i=1}^{K} |E_i(\mathbf{H}(\mathbf{x}_{c,t}))| \qquad \mathbf{H}^*(\mathbf{x}_{c,t}) = \arg\min(\text{MAE}(\mathbf{H}(\mathbf{x}_{c,t}))) \tag{3.34}$$

With this approach, we can use only the information of a localized part of the entropy-GDP plane, going beyond a single available time window forecast. Indeed, one close analogue to France in the year 2014 is United Kingdom in the year 2005, and to study its evolution, we can use any analogues up to 2009 (so their 5 year evolution is known at the test year 2014). The procedure to find the best bandwidth is the following.

1. We choose a point in the entropy-GDP plane $\mathbf{x}_{c,t}$ (where c is a country and t is the year) and find the K close analogues to it $\mathbf{x}_{\tilde{c},\tau}$ with $\tau \leq t - \Delta t$, where $\Delta t = 5$ years.
2. Using the set of K close analogues $\mathbf{x}_{\tilde{c},\tau}$ as a training set, we solve the optimization problem finding the best parameters that minimize the mean absolute error, using the entire datasets available with evolution know at the test year t.

In our approach we have considered $K = 10$, we did not find any considerable difference using other values. A few examples of the heterogeneity of bandwidth that we get using this algorithm can be found at the end of the chapter in Figure 3.7 (different income nations).

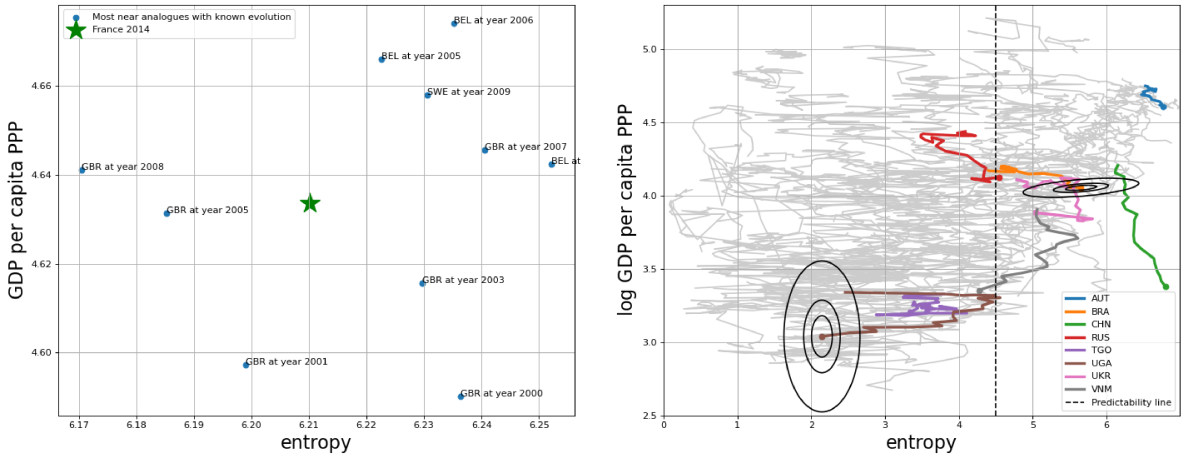


Figure. 3.4: On the left, we have the 10 most close analogues, on which their evolution is known at the test time, to France in the year 2014. In the right graph, we have an example of the best bandwidth learned investigating the known evolution from 2009 to 2014. The three ellipses are the locus of the point of constant probabilities of 0.8, 0.5 and 0.05; after the outermost line, no more analogues are considered.

3.4 Minimizing the Test MAE

In this part, we will study which is the best way to infer on a training set the bandwidth $\mathbf{H}$ that minimize the MAE of the test set evolutions from 2014 to 2019. We will study the cSPS using univariate (one parameter bandwidth) and multivariate (three parameters bandwidth) Gaussian kernel.
We computed the best bandwidth from an in-sample analysis of the growth from 2014 to 2019, hence by a direct minimization of the test set.

- one parameter: $\sigma^* = 0.24 \longrightarrow \text{MAE}(\sigma^*) = 1.39\%$
- three parameters: $(\sigma_x^*, \sigma_y^*, \theta^*) = (0.48, 0.2, 0) \longrightarrow \text{MAE}(\sigma_x^*, \sigma_y^*, \theta^*) = 1.38\%$

These are the optimal bandwidths, and so the highest accuracy, that we will try to infer using an out-of-sample analysis. All the MAE listed in the following will refer to the test set and we will not report the MAE from the training set.

Analysis with constant $\mathbf{H}^*$ We present the best bandwidths obtained by fitting 98 countries' known evolution from 2009 to 2014.

- one parameter: $\sigma^* = 0.11 \rightarrow \text{MAE}(\sigma^*) = 1.52\%$
- three parameters: $(\sigma_x^*, \sigma_y^*, \theta^*) = (0.29, 0.03, -0.04\,\text{rad}) \rightarrow \text{MAE}(\sigma_x^*, \sigma_y^*, \theta^*) = 1.54\%$

There is a better accuracy using one parameter. To improve the analysis, we can use the distinction between predictability and unpredictability region found in the second chapter (we have predictability, high collinearity, with an entropy $H_c >= 4.5$)[2]. We can specialize the bandwidths in those regions

[2]Whether a nation belongs to a region depends on its initial entropy.

- one parameter: for a total MAE = 1.53%

$$\sigma^*_{\text{unpred}} = 0.21 \rightarrow \text{MAE}(\sigma^*_{\text{unpred}}) = 1.44\%$$
$$\sigma^*_{\text{pred}} = 0.06 \rightarrow \text{MAE}(\sigma^*_{\text{pred}}) = 1.73\%$$

- three parameters: for a total MAE = 1.49%

$$\mathbf{H}^*_{\text{unpred}} = (0.23, 0.07, -0.2\,\text{rad}) \rightarrow \text{MAE}(\mathbf{H}^*_{\text{unpred}}) = 1.54\%$$
$$\mathbf{H}^*_{\text{pred}} = (0.28, 0.02, -0.04\,\text{rad}) \rightarrow \text{MAE}(\mathbf{H}^*_{\text{pred}}) = 1.40\%$$

The distinction in two different regions of predictability emerges from the different accuracy. The three parameters model captures better the proper distance among countries in the predictability region, but is too specialized for countries showing poor laminar dynamics, giving space to the one-parameter model in this region. A combined model that uses one parameter for the unpredictability region and three parameters for the rest returns a $\text{MAE}(\sigma^*_{\text{unpred}}, \mathbf{H}^*_{\text{pred}}) = 1.42\%$.
With this combined model, univariate cSPS for the unpredictability region and multivariate for the predictability one, we obtain the best out-of-sample predictions using the constant $\mathbf{H}^*$ algorithm. We call this statistical learning algorithm *combined constant* $\mathbf{H}^*$ (see in the right plot of the figure 3.4).

An in-sample minimization of the MAE in the test set using this new method returns a MAE = 1.37%. Therefore, using the combined constant $\mathbf{H}^*$ algorithm on the training set, we obtain bandwidths producing a MAE in the test set very close to the absolute minimum.

Analysis with specialized $\mathbf{H}^*(\mathbf{x}_{c,t})$ In this case, the best bandwidths are inferred from each country. Using $K = 10$ (the number of closed analogues to use for fitting the parameters) we obtained these results

- one parameter: for a total MAE = 1.56%

$$\text{MAE}(\{\sigma^*(\mathbf{x}_{c,t})\}_{\text{unpred}}) = 1.48\%$$
$$\text{MAE}(\{\sigma^*(\mathbf{x}_{c,t})\}_{\text{pred}}) = 1.64\%$$

- three parameters: for a total MAE = 1.51%

$$\text{MAE}(\{\mathbf{H}^*(\mathbf{x}_{c,t})\}_{\text{unpred}}) = 1.56\%$$
$$\text{MAE}(\{\mathbf{H}^*(\mathbf{x}_{c,t})\}_{\text{pred}}) = 1.51\%$$

With the one-parameter model, we get worst results than considering a constant parameter in the whole plane. We get worst results also if we use the three parameters model. This reduction in accuracy using a more complex model is a consequence of a localized over-fitting on the ten most close countries.

From these considerations, It is better to use a univariate cSPS in the unpredictability region and a multivariate cSPS in the predictability one. The bandwidth of the cSPS will be learn by minimizing the MAE of a training set (the most recent known evolutions at the test time) using the combined constant $\mathbf{H}^*$ algorithm. A recap of the method used and their accuracy in the test MAE can be found in Figure 3.5.

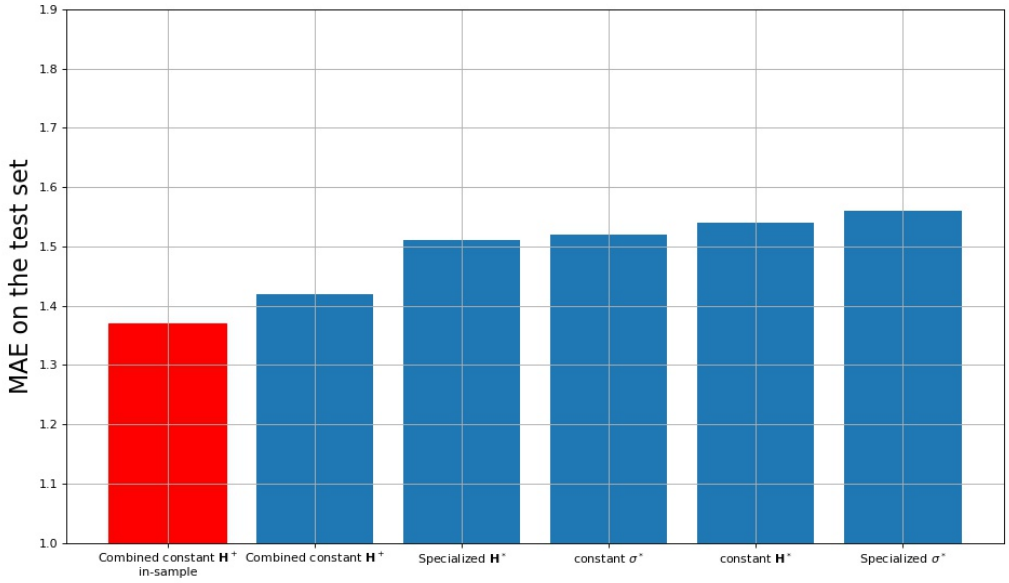


Figure. 3.5: Different accuracy in the test set (evolution from 2014 to 2019). In red we have the best possible accuracy obtained by a direct minimization in the test set using the combined constant $\mathbf{H}^*$, while in blue we have the test MAE obtained from a minimization in a training set. The best algorithm is the combined constant $\mathbf{H}^*$ while the others perform almost equally.

3.5 Self-Correlation of Growth: Velocity

Following the refinement of SPSb introduced in [36] we discuss the self-correlation of growth. With the cSPS approach, we sample analogues in a temporal range for which their evolution is known when we perform the forecasts, losing the information embedded into the recent past of the considered country's time series. Nonetheless, it is well known that GDP growth exhibit a strong self-correlation [54]. We perform a naive analysis of growth, that is that the country should move in the plane as much as it moved in the previous year, on average. The algorithm is the following

1. For each time series $\{\mathbf{x}_{c,t}\}_{t=T_0,\dots,T}$ we take only its Δt year history $\{\mathbf{x}_{c,t}\}_{t=T-\Delta t,T}$, for $\Delta t = 5$. We do not want to consider jumps made during situations in the past that could have been different from the current situation, so a time series with only historical data of 5 years in the past should be sufficient to estimate an auto-correlated growth.

2. We take for each year the one year velocities $\mathbf{v}_{c,t} = \mathbf{x}_{c,t+1} - \mathbf{x}_{c,t}$ then we compute the one year average velocity and its variance

$$\langle \mathbf{v}_c \rangle = \frac{1}{\Delta t} \sum_{t=T-\Delta t}^{T-1} \mathbf{v}_{c,t} \qquad \sigma^2_{\langle \mathbf{v}_c \rangle} = \frac{1}{\Delta t} \sum_{t=T-\Delta t}^{T-1} \left(\mathbf{v}_{c,t} - \langle \mathbf{v}_c \rangle \right)^2 \tag{3.35}$$

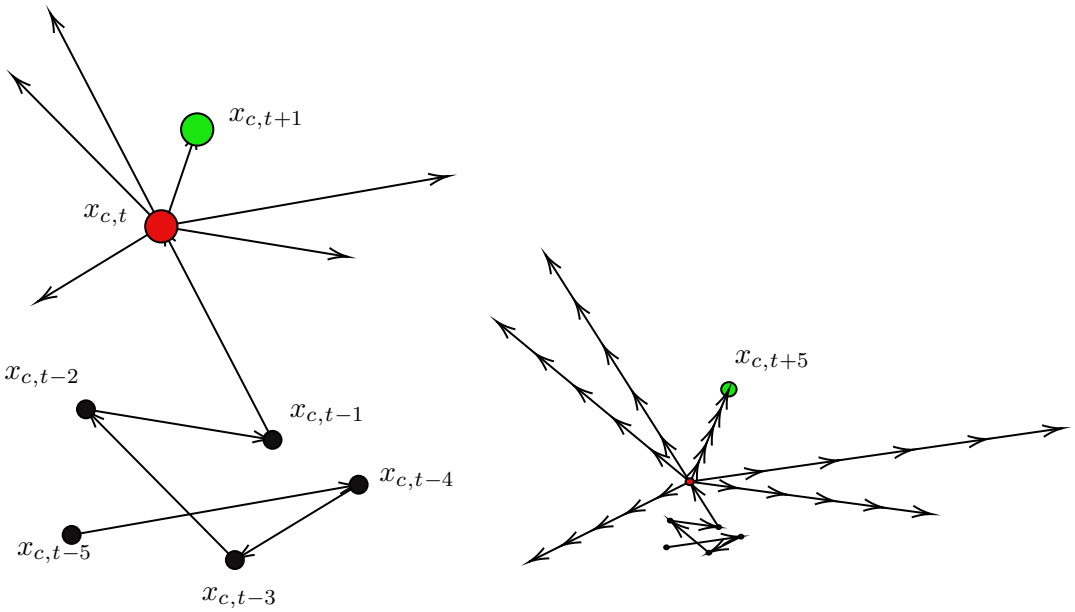


Figure. 3.6: Example of velocity algorithm

3. We assume that the mean 5 year evolution is 5 times the mean one year evolution, hence:

$$\langle \delta \mathbf{x}_{c,t} \rangle = \Delta t \langle \mathbf{v}_c \rangle = \sum_{\tau=T-\Delta t}^{T-1} \mathbf{v}_{c,t} \tag{3.36}$$

$$\sigma^2_{\langle \delta \mathbf{x}_{c,t} \rangle} = (\Delta t)^2 \sigma^2_{\langle \mathbf{v}_c \rangle} \tag{3.37}$$

This algorithm can be added to the cSPS providing a better accurate prediction.

3.5.1 Velocity-cSPS

We can incorporate the auto correlation information to the cSPS algorithm. We have two estimates of $\langle \delta \mathbf{x}_{c,t} \rangle$: one given by the cSPS algorithm $\langle \delta \mathbf{x}_{c,t} \rangle_{\text{cSPS}}$ and one given by the velocity algorithm $\langle \delta \mathbf{x}_{c,t} \rangle_{\text{vel}}$, each estimate has its own error. A new forecast that uses both methods is a weighted average of them.

$$\langle \delta \mathbf{x}_{c,t} \rangle = \alpha(c,t) \langle \delta \mathbf{x}_{c,t} \rangle_{\text{cSPS}} + (1 - \alpha(c,t)) \langle \delta \mathbf{x}_{c,t} \rangle_{\text{vel}} \tag{3.38}$$

For the univariate case, the parameter α is chosen by minimizing the variance of the new estimate. Let us consider two model A and B that give two estimates x_A and x_B with variance $\text{Var}[x]_A$ and $\text{Var}[x]_B$. The best weighting parameter α can be computed as

$$\text{Var}[x, \alpha] = \alpha^2 \text{Var}[x]_A + (1-\alpha)^2 \text{Var}[x]_B \tag{3.39}$$

$$\frac{\partial \text{Var}[x,\alpha]}{\partial \alpha} = 0 \quad \Longleftrightarrow \quad \alpha = \frac{\text{Var}[x]_B}{\text{Var}[x]_B + \text{Var}[x]_A} \tag{3.40}$$

For the multivariate case, the things are more subtle because we have to decide which component we want to minimize of the variance matrix. An idea would be to minimize the generalized variance, given by the determinant of the covariance matrix; however, this approach is not correct for our purpose as we have optimized the parameters of the cSPS by minimizing the error on the y-axis, regardless of the error in entropy evolution. So the right way is to minimize

the yy component of the variance matrix, getting in this sense a new estimate with the small possible variance in GDP.

$$\alpha(c,t) = \frac{\sigma^2_{yy}(\langle \delta \mathbf{x}_{c,t} \rangle_{\text{vel}})}{\sigma^2_{yy}(\langle \delta \mathbf{x}_{c,t} \rangle_{\text{vel}}) + \sigma^2_{yy}(\langle \delta \mathbf{x}_{c,t} \rangle_{\text{cSPS}})} \tag{3.41}$$

3.6 Average Improvement in Accuracy

We performed a back-test analysis on three different time windows: 2012-2017, 2013-2018 and 2014-2019, and computed the mean absolute error (MAE) on the overall predictions on a total of 275 observations to match the prediction made by the IMF. In addition, we will also report the root mean square error (RMSE) as a complementary accuracy measure of MAE.
We used the evolution from 2014 to 2019 as a test set to decide which is the best algorithm to infer the best bandwidth $\mathbf{H}$ in section 3.4, therefore the results presented here are not strictly out-of-sample for this window. However, the evolutions from 2012 to 2017 and from 2013 to 2018 can be considered as fully out-of-sampled, as the algorithm to infer $\mathbf{H}^*$ is already set.
The out-of-sample bandwidth selection optimization returned these values.

- Using as training set the evolution from 2007 to 2012:

$$\sigma^*_{\text{unpred}} = 0.12 \qquad \mathbf{H}^*_{\text{pred}} = (0.54, 0.02, -0.12\,\text{rad}) \qquad \text{MAE}_{2012\to2017} = 1.57\%$$

- Using as training set the evolution from 2008 to 2013:

$$\sigma^*_{\text{unpred}} = 0.14 \qquad \mathbf{H}^*_{\text{pred}} = (0.41, 0.01, -0.12\,\text{rad}) \qquad \text{MAE}_{2013\to2018} = 1.53\%$$

- Using as training set the evolution from 2009 to 2014:

$$\sigma^*_{\text{unpred}} = 0.21 \qquad \mathbf{H}^*_{\text{pred}} = (0.28, 0.02, -0.03\,\text{rad}) \qquad \text{MAE}_{2014\to2019} = 1.42\%$$

	MAE %	RMSE %	Improvement in MAE	Improvement in RMSE
IMF	1.88	2.37	0%	0%
velocity	2.07	2.87	−10%	−21%
cSPS with combined constant parameters σ^*_{unpred} and $\mathbf{H}^*_{\text{pred}}$				
cSPS	1.50	2.09	+20%	+14%
v-cSPS	1.46	2.02	+22%	+15%

Table. 3.1: Errors in 5 year prediction of GDP per capita PPP in constant 2017 international dollar for years 2017, 2018 and 2019, with a total of 275 predictions.

	Predictability ($H_c \geq 4.5$)	Unpredictability ($H_c < 4.5$)
v-cSPS MAE	1.28%	1.66%
v-cSPS RMSE	1.82%	2.21%

Table. 3.2: Difference in errors between predictability and unpredictability region. We have 143 observations in the predictability region, while 131 in the other one.

Using these bandwidths, we forecasted the five-year growth in GDP for 89 countries in 2012, 91 countries in 2013, and 95 countries for 2014. The results are listed in Table 3.1 in which also the velocity-cSPS algorithm (v-cSPS) is taken into account. The cSPS algorithm alone is responsible for an increment in the accuracy of the 20% regarding the IMF's predictions. The self-correlation of growth (velocity) is not a good model, but if we apply it to the cSPS we improve the accuracy for another 2%. The average value of the weighting parameter $\langle \alpha(c,t) \rangle = 0.49$ introduced in (3.38) indicates that the two algorithms contribute equally as they have similar errors.
The difference in errors between the two distinct region in the entropy-GDP plane justify their names: we have higher predictability in GDP growth in the predictability region.

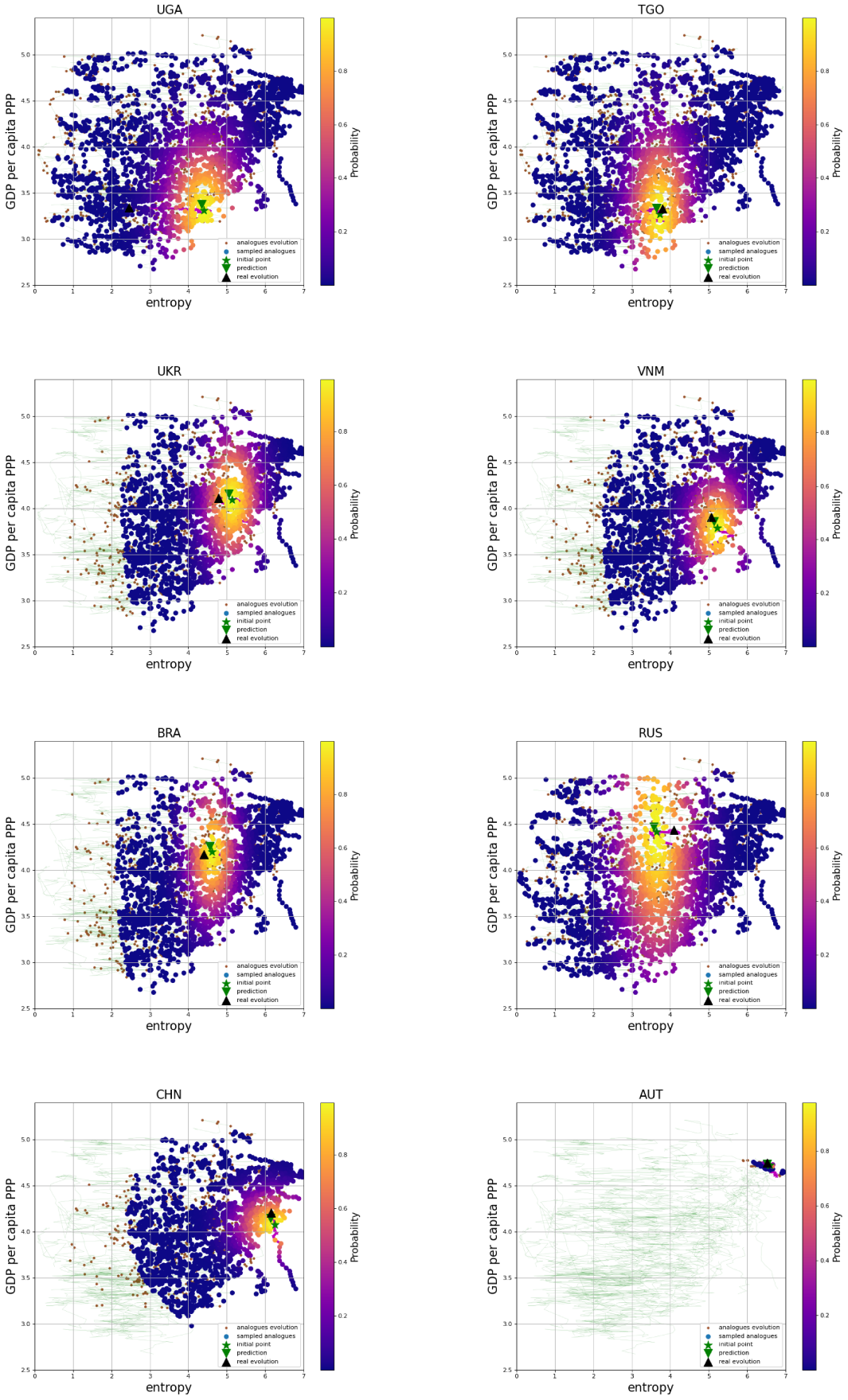


Figure. 3.7: Convergence selective predictability with specialized $\mathbf{H}^*$ algorithm applied to under, lower-middle, upper-middle and high developed nations. Estimates are from 2014 to 2019.

Conclusion

We discussed how diversification in exports has an impact on the wealth of nations, and we developed a new algorithm to measure it in the context of economic complexity. The new measure of nations' complexity and products' ubiquity fully use the information contained in the dataset, going beyond the theory of comparative advantage. A construction based on a self-consistent use of the Shannon entropy offers a measure with a universally accepted interpretation of non-trivial diversification of a nation's export that considers the different importance of products in world trade. In addition, the boundedness of the Shannon entropy gives a stable algorithm to compute the measures that exponentially converges to a unique fixed point both locally and globally.
Studying the year 2019 with the 2017 edition of HS, we showed that this new complexity measure positively correlates with income and income per capita, and it can qualitatively discriminate among nations and products according to the economic narrative. Furthermore, a new approach to measure intra and inter sectorial diversification emerged thanks to the coarse-grained property of the Shannon entropy, showing how true diversification is reached in any sector for the most complex nations. Moreover, thanks to this new algorithm, we can perform the sectorial aggregation still using the maximum information available at the finer level, going beyond a mere aggregation of products. Finally, the role of revealed comparative advantage has been discussed within this framework. We showed how the binarized version of the bi-adjacency matrix can lead one-product dependent countries to get high value of complexity, and how the products with low export share are neglected, especially for the less diversified nations.
We showed the heterogeneous dynamics of countries with a coarse-grained analysis on the entropy-GDPpcPPP plane describing the motion as if it would have a flow structure. A region of flow predictability emerged thanks to the entropy dimension, showing how high entropic countries have a more stable economy. We got similar results considering diversification in comparative advantage products (filtered RCA entropy) while a substantially different dynamics emerged focusing on just which products are exported without considering the amount of their exchange (binarized RCA entropy). Nonetheless, the emergent properties: predictability, negative correlation with entropy and entropy-variance (GDPpcPPP variance), and positive correlation with middle and long term growth, remain using these three different measures. This fact indicates that despite the debated role of RCA to unveil competitiveness, it filters the most crucial information of countries' exports. Full information entropy, which does not use RCA, has the advantage that no economic preconception is needed (comparative advantage is often criticized by economists who do not support neo-liberalism [55]), no choice of the RCA threshold and no data loss are required, and moreover, it still captures the essential information, making this approach a genuinely data-driven and stable against noise indicator.
We showed how the growth in GDPpcPPP can be described as a local trend in the entropy-GDPpcPPP plane, drastically diminishing the forecasting model's complexity offered by standard econometric approaches (IMF). We used a kernel regression version of the Selective Pre-

dictability Scheme bootstrap (SPSb), an algorithm based on the method of analogue, with a multivariate Gaussian kernel, calling it cSPS. In addition, we discussed and solved the problem to decide a priori the optimal parameters to identify analogues in the plane, using a statistical learning approach. Using as the test set the evolution from 2014 to 2019 and as the training set the most recent known evolution at the test time, hence from 2009 to 2014, we observed that a univariate cSPS works better in the unpredictability region, while a multivariate cSPS gives better result for the predictability one. With this approach, we found that we can outperform the accuracy of the IMF's prediction by twenty percent, using a less complex model and, more importantly, fewer data. Furthermore, the difference in errors between predictability and unpredictability corroborate that entropy discriminates between two different dynamical regimes.

The algorithm studied to forecast GDPpcPPP growth is general and it can be easily inserted in a generic problem of dynamical systems, on which the method of analogues can be used. Any complex system that can be described by a few number of variables is a candidate for this algorithm. The only requirement is that the dimension of the system (hence their number of variables) should be sufficient to individuate laminar dynamics, as shown in the second chapter. In addition, to properly work the algorithm needs a considerable number of historical data, especially with similar condition of the system we want to predicts.
The statistical learning algorithm developed allow us to test the algorithm on historical data before applying it. In addition, the statistical learning approach to infer the optimal bandwidth of the kernel function can be modified according to what we are interested to observe, by changing the minimization function.

Nadaraya-Watson kernel regression

Kernel regression is a non-parametric technique to estimate the conditional expectation of a random variable. Firstly introduced for univariate random variables we now present the theory for the multivariate case.
A probability density function of an ensamble of random vector i.i.d. $\{\mathbf{x}_i\}_{i=1,\dots,N}$ can be estimate as a sum of p.d.f. centered on those points, this technique is called *kernel density estimation.*

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i) \tag{A.1}$$

The p.d.f $K_{\mathbf{H}}$ is called kernel, which is a symmetric multivariate density function. The accuracy does not depend on the form of the kernel function, but it does on the choice of the smoothing parameter $\mathbf{H}$ that is called bandwidth of the kernel function.

$$\mathbf{H} \quad \text{Is a symmetric and positive definite matrix} \tag{A.2}$$

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}) \tag{A.3}$$

Since the accuracy of the estimation does not depend on the form of the kernel it is convenient to consider a multivariate Gaussian kernel

$$K_{\mathbf{H}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{|\mathbf{H}|}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{H}^{-1}\mathbf{x}} \tag{A.4}$$

where d is the dimension of the random vector $\mathbf{x}$. In this context the bandwidth plays the role of a covariance matrix.

Let's consider now a set of observation $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,N}$. We want to construct a non linear regression model of the form

$$\mathbf{y}_i = m(\mathbf{x}_i) + \epsilon_i \tag{A.5}$$

where ϵ_i is an error with zero mean. The best estimate of $\mathbf{y}$ given an observation $\mathbf{x}$ is the following

$$E(\mathbf{y}|\mathbf{x}) = \int \mathbf{y} f(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y} = \int \mathbf{y} \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})}\, \mathrm{d}\mathbf{y} \tag{A.6}$$

We can approximate the probability distributions using a kernel density estimation

$$f(\mathbf{x},\mathbf{y}) \simeq \frac{1}{N}\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i) K_{\mathbf{H}'}(\mathbf{y}-\mathbf{y}_i) \tag{A.7}$$

$$f(\mathbf{x}) \simeq \frac{1}{N}\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i) \tag{A.8}$$

where in principle we can use a different bandwidth for the disjoint set $\{\mathbf{y}_i\}_{i=1,\dots,N}$. Substituting in the estimate we get

$$E(\mathbf{y}|\mathbf{x}) \simeq \int \frac{\mathbf{y}\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)K_{\mathbf{H}'}(\mathbf{y}-\mathbf{y}_i)}{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)}\mathrm{d}\mathbf{y} \tag{A.9}$$

$$= \frac{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)\int \mathbf{y}K_{\mathbf{H}'}(\mathbf{y}-\mathbf{y}_i)\,\mathrm{d}\mathbf{y}}{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)} \tag{A.10}$$

$$= \frac{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)\mathbf{y}_i}{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)} = \hat{E}(\mathbf{y}|\mathbf{x}) \tag{A.11}$$

Where in the last step we used the fact that the kernel function $K_{\mathbf{H}'}(\mathbf{y}-\mathbf{y}_i)$ has mean equal to $\mathbf{y}_i$. The estimator found above is the so called Nadaraya-Watson estimator.

$$\hat{E}(\mathbf{y}|\mathbf{x}) = \hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x_i})\mathbf{y}_i}{\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x_i})} \tag{A.12}$$

For a multivariate Gaussian kernel this estimator reduces to

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\sum_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_i\right)^T \mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}_i\right)\right]\mathbf{y}_i}{\sum_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_i\right)^T \mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}_i\right)\right]} \tag{A.13}$$

Errors made by the IMF

Two times per year, the International Monetary Fund publishes the World Economic Outlook (WEO), where its staff of economists presents analyses of global economic developments during the near and medium-term.Despite the quality of these forecasts being debated in the economic literature [56], they remain a valid reference to investigate the accuracy of our algorithm.

Year	MAE %	RMSE %	Number of countries
2007-2012	$3.68^{+0.38}_{-0.37}$	$4.16^{+0.40}_{-0.37}$	100
2012-2017	$1.68^{+0.30}_{-0.25}$	$2.13^{+0.39}_{-0.29}$	89
2013-2018	$2.43^{+0.37}_{-0.32}$	$2.94^{+0.46}_{-0.34}$	91
2014-2019	$1.55^{+0.26}_{-0.21}$	$1.94^{+0.37}_{-0.27}$	95
TOTAL	$1.88^{+0.16}_{-0.26}$	$2.37^{+0.20}_{-0.25}$	275

Table. B.1: Errors in 5 year prediction of GDP per capita PPP in constant 2017 international dollar for years 2012, 2017, 2018 and 2019. The errors are computed with a bootstrap technique at confidence level 0.95. In the last row we have an aggregated errors from the three different time windows considered (2017, 2018 and 2019). To notice the high errors for the predictions from 2007 to 2012 as a consequence of the Great Recession.

Scaling Our approach makes use as a monetary unit the GDPpcPPP in constant 2017 international dollars; the PPP (purchasing power parity) is an index that allows comparing better prices in different locations, while we used constant dollars to avoid inflation growth. WEO reports countries listed by GDPpcPPP, but in the current international dollar, an exchange is therefore required.
From the World Bank, we can acquire the databases of GDPpcPPP in current international dollars (we call this database "current_{df}") and of GDPpcPPP in constant 2017 international dollars (we call this database "constant_{df}"), the one we have used for our analysis. We can compute an exchange rate for each county at each year using these two databases as a ratio

$$\alpha(c,t) = \frac{\text{constant}_{df}(c,t)}{\text{current}_{df}(c,t)} \tag{B.1}$$

Where $\text{current}_{df}(c,t)$ ($\text{constant}_{df}(c,t)$) indicates the GDPpcPPP in current (constant) international dollars of the country c at the year t. We can now apply these exchange rates to the IMF database

$$\big(\text{IMF}(c,t), \text{IMF}(c,t+\Delta t)\big) \mapsto \big(\alpha(c,t)\text{IMF}(c,t), \alpha(c,t+\Delta)\text{IMF}(c,t+\Delta t)\big) \tag{B.2}$$

Predictions We gather all the released WEO second semester reports from 2007 to 2019 in an unique dataset and extract the the five year prediction in GDPpcPPP in constant dollar from 2017 to 2019. These reports are corrected in a regular basis [1] so we can find different forecasts and also different initial GDPpcPPP values for the same country. To extract an unique information we average the repeated values getting in this way an unique forecasts. After these process we perform the scaling and then we compute the predicted compound annual growth rate.
To compute the error we used the World Bank dataset of GDPpcPPP in constant 2017 international dollars as reference (the same used in the entropy-GDPpcPPP plane and to compute the scaling factor for the IMF database). The predictions are then an intersection of countries listed in IMF and World Bank dataframe. In table B.1 we collect the mean absolute errore (MAE) and root mean square error (RMSE) for each year.

[1] Historical data are updated on a continual basis as more information becomes available, and structural breaks in data are often adjusted to produce smooth series with the use of splicing and other techniques [57]

Bibliography

[1] Samir Suweis et al. "Emergence of structural and dynamical properties of ecological mutualistic networks". In: *Nature* 500.7463 (2013), pp. 449–452.

[2] Danielle S Bassett, David L Alderson, and Jean M Carlson. "Collective decision dynamics in the presence of external drivers". In: *Physical Review E* 86.3 (2012), p. 036105.

[3] Filippo Simini et al. "A universal model for mobility and migration patterns". In: *Nature* 484.7392 (2012), pp. 96–100.

[4] Robert K Adair. "Stochastic contributions to global temperature changes". In: *Physical review letters* 100.14 (2008), p. 148501.

[5] Victor M Eguiluz et al. "Scale-free brain functional networks". In: *Physical review letters* 94.1 (2005), p. 018102.

[6] Pierre-Alexandre Balland et al. "The new paradigm of economic complexity". In: *Research Policy* 51.3 (2022), p. 104450.

[7] Adam Smith. *The wealth of nations [1776]*. Vol. 11937. na, 1937.

[8] David Ricardo. *On the principles of political economy*. J. Murray London, 1821.

[9] Opportunity cost. *Opportunity cost — Wikipedia, The Free Encyclopedia*. [Online; accessed 14-Februar-2022]. 2022. URL: `https://en.wikipedia.org/wiki/Opportunity_cost`.

[10] Eli F Heckscher. "Utrikeshandelns verkan på inkomstfördelningen. Några teoretiska grundlinjer". In: *Ekonomisk tidskrift* (1919), pp. 1–32.

[11] Bertil Ohlin. "Till frågan om penningteoriens uppläggning". In: *Ekonomisk Tidskrift* (1933), pp. 45–81.

[12] Paul R Krugman. "Increasing returns, monopolistic competition, and international trade". In: *Journal of international Economics* 9.4 (1979), pp. 469–479.

[13] Edward E Leamer and James Levinsohn. "International trade theory: the evidence". In: *Handbook of international economics* 3 (1995), pp. 1339–1394.

[14] Ma Angeles Serrano and Marián Boguná. "Topology of the world trade web". In: *Physical Review E* 68.1 (2003), p. 015101.

[15] César A Hidalgo et al. "The product space conditions the development of nations". In: *Science* 317.5837 (2007), pp. 482–487.

[16] Michele Caraglio, Fulvio Baldovin, and Attilio L Stella. "Export dynamics as an optimal growth problem in the network of global economy". In: *Scientific reports* 6.1 (2016), pp. 1–10.

[17] Gianluca Teza, Michele Caraglio, and Attilio L Stella. "Data driven approach to the dynamics of import and export of g7 countries". In: *Entropy* 20.10 (2018), p. 735.

[18] Gianluca Teza, Michele Caraglio, and Attilio L Stella. "Growth dynamics and complexity of national economies in the global trade network". In: *Scientific reports* 8.1 (2018), pp. 1–8.

[19] Jan Tinbergen. "Shaping the economy; suggestions for an international economic policy". In: (1962).

[20] Jean-Philippe Bouchaud and Marc Mézard. "Wealth condensation in a simple model of economy". In: *Physica A: Statistical Mechanics and its Applications* 282.3-4 (2000), pp. 536–545.

[21] Mehran Kardar, Giorgio Parisi, and Yi-Cheng Zhang. "Dynamic scaling of growing interfaces". In: *Physical Review Letters* 56.9 (1986), p. 889.

[22] Harrison Searles. "César Hidalgo: Why information grows: The evolution of order, from atoms to economies". In: *The Review of Austrian Economics* 30 (Sept. 2015). DOI: 10.1007/s11138-015-0328-6.

[23] Ricardo Hausmann and César A Hidalgo. "The network structure of economic output". In: *Journal of Economic Growth* 16.4 (2011), pp. 309–342.

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

[25] Luca De Benedictis and Massimo Tamberi. "A note on the Balassa index of revealed comparative advantage". In: *Available at SSRN 289602* (2001).

[26] Bela Balassa. "Trade liberalisation and "revealed" comparative advantage 1". In: *The manchester school* 33.2 (1965), pp. 99–123.

[27] Penny Mealy and Alexander Teytelboym. "Economic complexity and the green economy". In: *Research Policy* (2020), p. 103948.

[28] Dominik Hartmann et al. "Linking economic complexity, institutions, and income inequality". In: *World development* 93 (2017), pp. 75–93.

[29] Trung V Vu. "Economic complexity and health outcomes: A global perspective". In: *Social Science & Medicine* 265 (2020), p. 113480.

[30] Eric Kemp-Benedict. "An interpretation and critique of the Method of Reflections". In: (2014).

[31] Andrea Tacchella et al. "A new metrics for countries' fitness and products' complexity". In: *Scientific reports* 2.1 (2012), pp. 1–7.

[32] Greg Morrison et al. "On economic complexity and the fitness of nations". In: *Scientific reports* 7.1 (2017), pp. 1–11.

[33] Orazio Angelini et al. "The complex dynamics of products and its asymptotic properties". In: *PloS one* 12.5 (2017), e0177360.

[34] Orazio Angelini and Tiziana Di Matteo. "Complexity of products: the effect of data regularisation". In: *Entropy* 20.11 (2018), p. 814.

[35] Matthieu Cristelli, Andrea Tacchella, and Luciano Pietronero. "The heterogeneous dynamics of economic complexity". In: *PloS one* 10.2 (2015), e0117174.

[36] Andrea Tacchella, Dario Mazzilli, and Luciano Pietronero. "A dynamical systems approach to gross domestic product forecasting". In: *Nature Physics* 14.8 (2018), pp. 861–865.

[37] Edward N Lorenz. "Atmospheric predictability as revealed by naturally occurring analogues". In: *Journal of Atmospheric Sciences* 26.4 (1969), pp. 636–646.

[38] Charles D Brummitt et al. "Machine-learned patterns suggest that diversification drives economic development". In: *Journal of the Royal Society Interface* 17.162 (2020), p. 20190283.

[39] Gianluca Teza, Michele Caraglio, and Attilio L Stella. "Entropic measure unveils country competitiveness and product specialization in the World trade web". In: *Scientific Reports* 11.1 (2021), pp. 1–11.

[40] World Customs Organization. *Harmonised System Nomenclature 2017 Edition*. 2017. URL: http://www.wcoomd.org/en/topics/nomenclature/instrument-and-tools/hs-nomenclature-2017-edition/hs-nomenclature-2017-edition.aspx.

[41] Zhuo-Ming Ren, Xiao Pan, and Yi-Cheng Zhang. "Significance of the Nested Structure in Multiplex World Trade Networks". In: *Complexity* 2020 (2020).

[42] Ricardo Hausmann et al. *The Atlas of Economic Complexity: Mapping Paths to Prosperity. Cambridge: Center for International Development, Harvard University.* 2011.

[43] Ricardo Hausmann. "The specialization myth". In: *Social Europe Journal* (2013).

[44] World Bank: Data. *Country classification.* 2022. URL: `https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups`.

[45] R Bruce Kellogg, Tien-Yien Li, and James Yorke. "A constructive proof of the Brouwer fixed-point theorem and computational results". In: *SIAM Journal on Numerical Analysis* 13.4 (1976), pp. 473–483.

[46] Lawrence C Evans and Ronald F Garzepy. *Measure theory and fine properties of functions.* Routledge, 2018.

[47] World Bank: Data. *Population Tot.* 2022. URL: `https://data.worldbank.org/indicator/SP.POP.TOTL`.

[48] World Bank: Data. *GDP (current USD).* 2022. URL: `https://data.worldbank.org/indicator/NY.GDP.MKTP.CD`.

[49] World Bank: Data. *GDP per capita (current USD).* 2022. URL: `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`.

[50] Joseph E Stiglitz, H Marcus, DM Hawley, et al. "GDP is the wrong tool for measuring what matters". In: *Scientific American* 1 (2020).

[51] Robert V Hogg and Allen T Craig. "Introduction to mathematical statistics.(5"" edition)". In: *Englewood Hills, New Jersey* (1995).

[52] CEPII. *Description of BACI.* 2017. URL: `http://www.cepii.fr/DATA_DOWNLOAD/baci/doc/DescriptionBACI.html`.

[53] Michaeël Bensimhoun. "N-dimensional cumulative function, and other useful facts about gaussians and normal densities". In: *Jerusalem, Israel, Tech. Rep* (2009), pp. 1–8.

[54] Lant Pritchett and Lawrence H Summers. *Asiaphoria meets regression to the mean.* Tech. rep. National Bureau of Economic Research, 2014.

[55] James K Galbraith. *The predator state: How conservatives abandoned the free market and why liberals should too.* Simon and Schuster, 2008.

[56] Axel Dreher, Silvia Marchesi, and James Raymond Vreeland. "The political economy of IMF forecasts". In: *Public Choice* 137.1 (2008), pp. 145–171.

[57] International Monetary Fund: World Economic Outlook Database. *Assumptions and Data Conventions.* 2021. URL: `https://www.imf.org/external/pubs/ft/weo/data/assump.htm`.