Journal of Informetrics 17 (2023) 101357

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

# Geography of science: Competitiveness and inequality

Aurelio Patelli[a,e,*], Lorenzo Napolitano[b], Giulio Cimini[c,e,a], Andrea Gabrielli[d,e,a]

[a] Enrico Fermi Center for Study and Research, via Panisperna 89a, 00184, Rome Italy
[b] European Commission, Joint Research Centre (JRC Seville), calle Inca Garcilaso 3, 41092 Seville Spain
[c] Physics Department and INFN, University of Rome Tor Vergata, Rome 00133 Italy
[d] Engineering Department, University "Roma Tre", 00146 Rome Italy
[e] Institute for Complex Systems (CNR), UoS Sapienza, 00185 Rome Italy

ARTICLE INFO

*Keywords:*
Science of science
Fitness and complexity
Scientific competitiveness
Inequality

ABSTRACT

We characterize the temporal dynamics of Scientific Fitness, as defined by the Economic Fitness and Complexity (EFC) framework, and R&D expenditures at the geographic scale of nations. Our analysis highlights common patterns across similar research systems, and shows how developing nations (China in particular) are quickly catching up with the developed world. This paints the picture of a general growth of scientific and technical capabilities of nations induced by the spreading of information typical of the scientific environment. Shifting the focus of the analysis to the regional level, we find that even developed nations display a considerable level of inequality in the Scientific Fitness of their internal regions. Further, we assess comparatively how the competitiveness of each geographic region is distributed over the spectrum of research sectors. Overall, the Scientific Fitness represents the first high quality estimation of the scientific strength of nations and regions, opening new policy-making applications for better allocating resources, filling inequality gaps and ultimately promoting innovation.

## 1. Introduction

Science is based on the progressive augmentation of existing knowledge building on past discoveries through a recursive process involving empirical observation and the formulation of testable hypotheses. Similarly to what happens for technological innovation and economic growth (Hidalgo & Hausmann, 2009; Pugliese et al., 2017; Tacchella et al., 2013), scientific progress requires appropriate capabilities: previous knowledge, tools, human capital, resources, and so on. The combination and interaction of such capabilities, even from different contexts, pushes the boundary of science through new knowledge and discoveries, as well as through re-discoveries via previously uncharted paths (Dosi et al., 2000; Iacopini et al., 2020; Tria et al., 2014). This process naturally occur mostly in geographic areas where many different capabilities are concentrated (Balland et al., 2020), whence we can assume that the scientific output of a region reflects the set of relevant capabilities available.

The quantitative evaluation of scientific outcomes, from the microscopic level of individual researchers and institutions to the macroscopic case of entire nations, is nowadays a common practice (Fortunato et al., 2018; Waltman, 2016). The seminal work by May (1997) assessed the performance of national research systems using an index borrowed from the economic literature: the Revealed Comparative Advantage (RCA) (Balassa, 1965), computed on the number of scientific documents produced by each nation in the various research sectors. King (2004) pursued a different approach, ranking nations according to the share of global citations received by their documents, and introduced funding as an additional variable of the analysis. Subsequently, the use of citations

[*] Corresponding author.
*E-mail address:* aurelio.patelli@cref.it (A. Gabrielli).

https://doi.org/10.1016/j.joi.2022.101357
Received 25 March 2022; Received in revised form 1 September 2022; Accepted 16 November 2022

became the gold standard for assessing research quality, and several metrics have been proposed, which measure different, albeit complementary, aspects connected to the quality of research outputs (Waltman, 2016). However, a purely citation-based approach has recently been questioned, due for example to the very different amount of resources that nations invest in scientific research. In fact, even for the most economically developed nations, scientific success as measured by citations and by public spending in Research and Development (R&D) (Cimini et al., 2016; Laverde-Rojas & Correa, 2019) are correlated but present strong deviations. Therefore, they should be considered as complementary dimensions for a correct evaluation of scientific performances.

There are two additional key aspects that citation share metrics do not take into account. On the macroscopic scale, nations do not specialize in a few research sectors, but tend to diversify their activity into as many sectors as possible. This is explained by the capability-based view, according to which a given geographic area is active in all research sectors allowed by the set of capabilities that it hosts. Since capabilities are heterogeneously distributed, nations have a heterogeneous level of diversification, thus diversification itself can be used as a proxy of scientific performance. Furthermore, while nations with many different capabilities (typically, the developed economies) are competitive in almost all existing research sectors, nations with fewer capabilities perform well only in a few research areas with a lower degree of sophistication. Such a *nested* structure, induced by the capability scheme, indicates the presence of a competitive mechanism shaping the connections between scientific actors – akin to what is observed in natural ecosystems (Mariani et al., 2019) or in human productive activities (Hidalgo & Hausmann, 2009). Indeed, although the scientific environment is neither directly nor indirectly aimed at the production of physical goods or services and it is not subject to the incentives of competitive markets, there are many sources of competition because most research systems rely on merit-based processes to determine funding, hiring, careers, and thus indirectly the direction of scientific research itself.

Overall, the nested pattern that emerges from the comparison of national research systems (Cimini et al., 2014) suggests that diversification and composition of the scientific research basket can be used to measure the scientific competitiveness (or Fitness) of a nation; at the same time, the Complexity of a research sector depends on its ubiquity and on the Scientific Fitness of nations that are competitive in that sector. The Economic Fitness and Complexity (EFC) algorithm (Cristelli et al., 2013; Cristelli et al., 2015; Tacchella et al., 2012) is the ideal tool to estimate the fixed point of this circular relation. The purpose of this work is precisely to implement a framework for quantifying scientific competitiveness by leveraging the EFC toolbox.

In a nutshell, we build an appropriate database for our analysis starting from the Open Academic Graph (OAG) (Färber, 2019; Sinha et al., 2015; Tang et al., 2008), a freely accessible collection of information about scientific publications. OAG assigns documents to geographic areas according to the location of the affiliation of the authors, and it assigns documents to research sectors according to a hierarchical classification of scientific topics, known as *Field of Studies* (FoS). The documents produced by a geographic area in a research sector provide a basic measure of scientific performance through an appropriate count of the number of received citations. Each paper is thus allocated to all its affiliations and FoS using fractional counting. In this analysis, we can vary the resolution both in terms of geographic scale (we follow the Territorial Level scheme implemented by OECD, 2020b and of FoS hierarchical level. Filtering this data using the RCA allows obtaining the Scientific Bipartite Network (SBN) connecting geographic areas with the research sectors they are competitive in, finally computing the Scientific Fitness (Cimini et al, 2014) of such areas through the EFC algorithm.

## 2. Data

We extract the scientific database from the Open Academic Graph (OAG)[1], a freely available snapshot of a two billion-scale academic graph resulting from the unification of Microsoft Academic Graph and AMiner (Färber, 2019; Sinha et al., 2015; Tang et al., 2008). We use OAG v2, created at the end of November 2018. The database records contributions to various kinds of outlets for scientific literature: journal articles, books, conference proceedings, reviews, and others. The coverage of OAG is estimated to be comparable to that of Scopus or WoS (Hug & Brändle, 2017), thus likely presenting similar geographic and phonetic biases – in particular the partial coverage of non-English written literature, especially in Social Sciences and Humanities, where research output is often published in the native language of the authors (Sivertsen & Larsen, 2012). We do not include in the analysis OAG entries related to patents, because we believe that patent documents are different from scientific publications especially in terms of the motivations leading to their production. OAG data spans more than a century, starting in principle at the beginning of 1800. In practice, data before the Second World War presents large fluctuations mainly due to the scarce amount of scientific production for most of the regions. Hence, we start the analysis in 1960, although the core results are presented only for the recent decades where also expenditure data is available (see below).

The classification of research sectors is defined by the Fields of Study (FoS), i.e. features that are dynamically evaluated by an "*in-house knowledge base related entity relationship, which is calculated based on the entity contents, hyperlinks and web-click signals*" (Tang et al., 2008). The FoS are mostly organized into a hierarchical structure, with the peculiarity that a code may have more parents.[2] This structure presents a static layer 0 with 19 hand-defined codes, corresponding to the main classification of the research sectors. Moving deeper in the hierarchy, layer 1 presents 294 codes while layer 2 has more than 80 000 codes and this number may change in time when new FoS are generated.[3]

[1] https://www.microsoft.com/en-us/research/project/open-academic-graph/.

[2] The very few exceptions of codes that are labelled at a fine level but without information on their parents are removed from the analysis. This does not represent a problem, since we consider only the highest levels of the FoS hierarchy.

[3] Deeper layers 3, 4 and 5 mostly split the larger topics, but are not considered in the present work.

The expenditure database is based on the available data collected by the OECD on the Gross Expenditure in Research and Developments (GERD) indices (OECD, 2020a). The database covers 48 countries, *i.e.* all the OECD members and few other relevant nations for which the data is made available, such as China and Russia. However, data quality depends strongly on national features, and the HERD database implemented in the analysis above is made available for 42 nations (among the OECD members only Colombia does not provide information of the expenditure). R&D expenditure is not the only reasonable proxy for input into R&D. For instance, Petersen et al. (2019) include the number of researchers employed in R&D activities among the elements contributing to the development of national scientific systems. Information of this kind would be greatly helpful to complement the present analysis and test its robustness. Unfortunately, the data at our disposal covers R&D employees only partially, and therefore does not allow us to expand the analysis in this direction with sufficient confidence. We hope to address this point in future work.[4] We implement a linear interpolation reconstructing the missing points. At TL2, the database follows the same classification implemented by the derivation of the territorial level SBN, edited by the OECD. However, the reconstruction at the finer geographic scales is interpolated keeping constant the national performances, since the data presents more than 50% of entries are missing in the data.

## 3. Methods

The OAG database is used to construct the bipartite network linking geographic areas to research sectors. To this end, we select only the OAG entries with full information about the authors' affiliation, FoS, citations count, and year of publication. Using this data we build tables reporting, for each year, the number of scientific documents produced by each geographic area in the various FoS, and the citations received up to the OAG creation date. In order to assign a document to a geographic area, OAG uses the location of the authors' main affiliation. Note that in some cases it is not possible to select a precise location because the affiliation may address generically a multinational firm or a multi-location research council (such as CNRS in France or CNR in Italy). In these cases the location of the headquarter is used, although this process may artificially boost the importance of capital regions. Note also that there are several documents associated with multiple FoS or author affiliations. In these cases, we employ a fractional counting approach by assigning the document to FoS and geographic areas with a weight that is inversely proportional to the number of FoS and number of authors.[5] Fractional counting has the main advantage that it allows aggregating tables along both the geographic and FoS dimensions without increasing disproportionately the weight of the most productive actors. Additionally, fractional countingis prefereble to other counting methods because it better balances the scientific outputs of large and small geographic regions (Aksnes et al., 2012).

Following the classical approach of the *Scientometrics* literature, we use citations received by scientific documents as a reliable proxy for the quality of research (Waltman, 2016). However, the simple citations count has some drawbacks, especially when used to assess a small corpus of papers. This is due to the time papers need to reach a stable level of citations (Burrel, 2002), to the high skewness of the citation distribution for single papers (Radicchi et al., 2008; Romeo et al., 2003), and to the dependence of citation patterns on the specific sector and journal considered. Indeed, the dynamical process underlying the evolution of citation counts is well modeled using a preferential attachment process (Eom & Fortunato, 2011; Medo et al., 2011; Wang et al., 2013). This means that the sum of the citations accrued by a set of papers is dominated by the citations of the few most cited outliers, which are in turn subject to strong statistical fluctuations (especially in small sets). A simple yet effective approach to reduce such fluctuations as well as the skewness of the citations distribution consists in using a logarithmic transformation (Fairclough & Thelwall, 2015; Medo & Cimini, 2016). Thus, we employ the *log-citations* count

$$w_{gs} = \log(1 + c_{gs}) \tag{1}$$

where $g$ labels a geographic area and $s$ a research sector, while $c_{gs}$ is the citation count of documents assigned to area $g$ and FoS $s$ published in a given year.

We further filter the log-citations counts to build a Scientific Bipartite Network (SBN), linking, for each year, the geographic areas with the research sectors in which they are competitive . To this end, we use an index borrowed from the economics literature, RCA, which proxies competitiveness with the ability of an actor to produce a higher output in a domain of activity than a reference level determined by the global average output in the same domain of activity. For our purposes, a geographic area is considered competitive in a research sector if its RCA is above a threshold value, typically set to 1. In formula:

$$\mathrm{RCA}_{gs} = \frac{w_{gs}}{\sum_{s'} w_{gs'}} \Bigg/ \frac{\sum_{g'} w_{g's}}{\sum_{g's'} w_{g's'}} \tag{2}$$

We thus build the SBN using the binary filter $M_{gs} = 1$ if $\mathrm{RCA}_{gs} \geq 1$ and $M_{gs} = 0$ otherwise. Note that, before implementing the filter, we apply an exponential smoothing to the RCA series, considering a half-life of 3 years in order to keep a short persistence in the data.

Finally, we feed the SBN to the EFC algorithm (Cristelli et al., 2013; Pugliese et al., 2016; Tacchella et al., 2012). The idea behind EFC is that the bipartite network connecting geographic areas to the outputs they produce in a domain of interest contains information about regional capabilities. For instance, the network that links countries to the products they export can tell us which

[4] We thank an anonymous referee for pointing out this issue.

[5] For example, if a paper is labeled with FoS $s_1$ and $s_2$ and has three authors, the first two affiliated with (also different) institutions in area $g_1$ and the third with an institution in area $g_2$, the paper is assigned to FoS $s_1$ and $s_2$ with the same weight 1/2, while it is assigned to geographic areas $g_1$ and $g_2$ respectively with weights 2/3 and 1/3. The paper's citations are split according to the same ratios.

products require the most advanced skills (i.e. have higher complexity), and which countries have the most advanced skill endowment (i.e. have higher fitness). Applying the EFC algorithm to scientific publications allows determining the Fitness of countries or regions and the Complexity of the fields in which they publish. The technical details of the algorithm can be found in (Pugliese et al., 2016; Tacchella et al., 2012). In a nutshell, the Fitness of a region is defined as the sum of the complexity of the fields to which it is linked, while the Complexity of a scientific field is mainly influenced by the country or region with the lowest Fitness actively producing scientific publication in it.

The EFC method exploits the nested structure of the network and computes the Fitness $F_g$ of a geographic area $g$ (which is a proxy for its global scientific competitiveness) by aggregating the complexities of its basket of research sectors in a non-linear way (so that the most complex sectors of activity weigh the most), and in the same way the complexity $Q_s$ of a research sector $s$ is given by the Fitness of the geographic areas that are active in it (with low competitive regions weighing the most). Operationally, the Fitness and the Complexity vectors are the fixed point of the following non-linear iterative map

$$\widetilde{F}_g^{(n)} = \sum_s M_{gs} Q_s^{(n-1)} \quad F_g^{(n)} = \frac{\widetilde{F}_g^{(n)}}{\langle \widetilde{F}_g^{(n)} \rangle_g}$$

$$\widetilde{Q}_s^{(n)} = \frac{1}{\sum_g M_{gs} \frac{1}{F_g^{(n)}}} \quad Q_s^{(n)} = \frac{\widetilde{Q}_s^{(n)}}{\langle \widetilde{Q}_s^{(n)} \rangle_s}$$

where the operator $\langle \cdot \rangle_x$ indicates the arithmetic mean with respect to the possible values taken by the variable dependent on the set $x$. Fixed point values of the Fitness are finally normalized by a reference value, which is taken to be the Fitness of the United States at TL1 aggregation and that of California (code US06) at TL2 aggregation. The normalization aims to regularize the heterogeneous distribution of Fitness across the years, thus highlighting the relative strength of the nations instead of a global competitiveness. Note that we build two kind of Fitness indicators: the Scientific Fitness based on log-citations counts of eq. (1), and the document Fitness based on log-document counts instead of citations.

We quantify the degree of scientific inequality within a nation using the Gini coefficient estimated on the dispersion of citation counts among its regions (Cozza & Schettino, 2015;in SI we consider a version of the Gini coefficient that takes population size into account). For our purposes, the Gini index can be written as follows:

$$G = 1 - \frac{\sum_{i+1}^{n} f(y_i)(S_{i-1} + S_i)}{S_n}$$

where $S_i = \sum_{j=1}^{i} f(y_j) y_j$, $S_0 = 0$, $f(y_i)$ is the fraction of regions within the same country that has received at least $y_i$ citations, and $y_i < y_j$ whenever $i < j$.

Recall that, in principle, the OAG database allows constructing the SBN at different levels of geographic aggregation, ranging from the fine-grained description of individual institutions to the macroscale of regions and nations. In this work, we focus on the macroscopic scale, in order to compare with previous literature about EFC and Science-of-Science. Leveraging the OECD Territorial Level Classification OECD, 2020b we generate the SBN both at two levels of aggregation: the Territorial Level 1 (TL1) comprising 207 countriesand the Territorial Level 2 (TL2) aggregation, which includes 577 regions[6] in 43 countries (some of which are not OECD members).

## 4. Results

### 4.1. National level analysis

We start by discussing Scientific Fitness at the geographic scale of nations – corresponding to the OECD Territorial Level 1 (TL1). Unsurprisingly, the geographic distribution of Fitness values, reported in the top map of Fig. 1, shows that the most developed nations in an Economic sense are also the top performers in Science, and the developing nations follow them. We observe a strong correlation with the Economic Fitness computed using export data (Cristelli et al., 2013; Cristelli et al., 2015; Tacchella et al., 2012), which measures competitiveness based on industrial capabilities, thought there is not a one-to-one correspondence between the two measures (Patelli et al., 2021; see SI for a further comparison). Notably, the ranking of Scientific Fitness is also different from that obtained using metrics based on citation shares, such as the Mean Normalized Citation Score (MNCS) Cimini et al. (2016); Waltman (2016), which measure research efficiency rather than competitiveness. Indeed, on tops of MNCS there are small but efficient research systems, such as Switzerland, Israel and Singapore. Instead, Scientific Fitness accounts both for efficiency (through the use of the RCA filter) and diversification (i.e., the cumulative stock of capabilities owned by a nation), and thus accounts for a fair comparison between small and large research systems. Remarkably, the same patterns are observed also when the analysis is performed using a different dataset (we report the case of Scimago ScimagoLab (2020) in the SI), supporting the consistency of our results.

In line with the previous literature on Science of Science (Abramo and D'Angelo, 2014; Cimini et al., 2016; King, 2004), we enrich the picture by contrasting Scientific Fitness with the amount of resources that are invested in scientific research. A similar

[6] There are in principle more than 700 regions but for some of them there is no affiliation found.

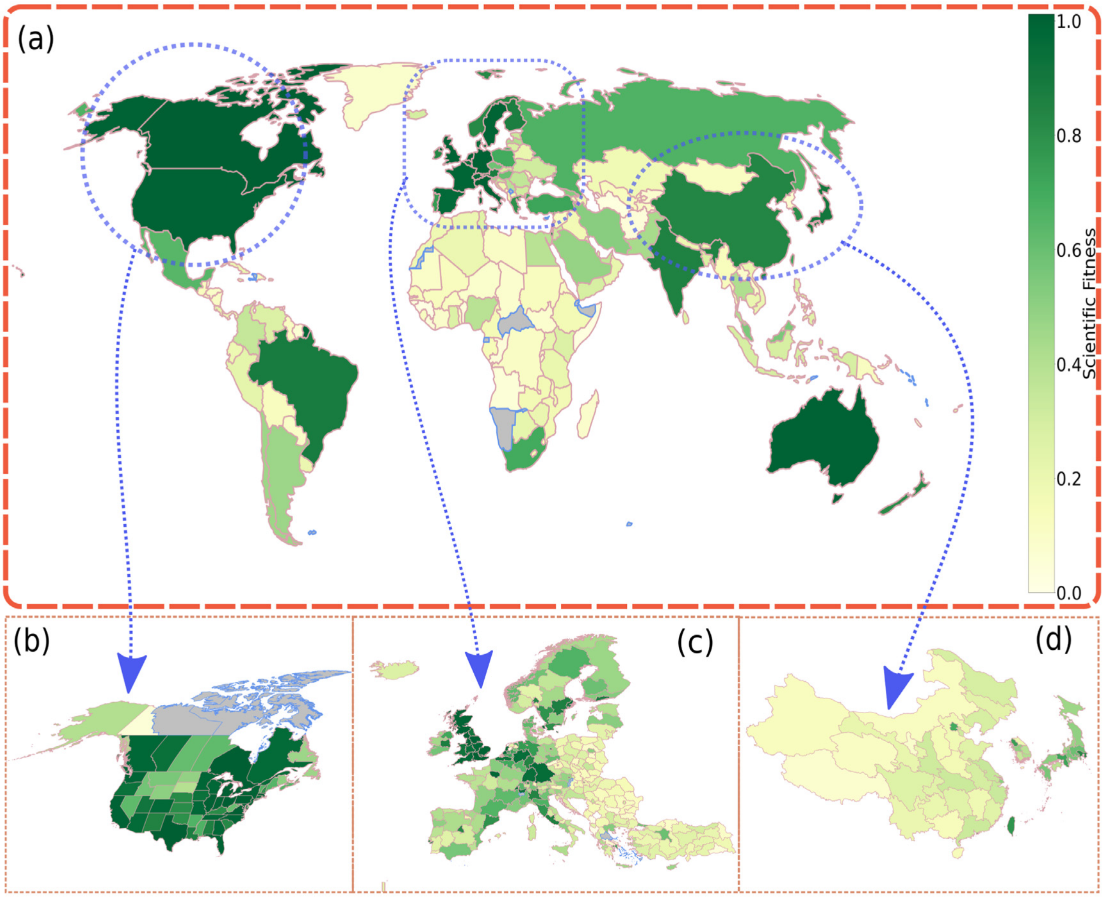

**Fig. 1.** Map of the Scientific Fitness of nations (TL1, panel a) and of regions (TL2) within North America (panel b), Europe and Turkey (panel c) and China, Japan and South Korea (panel d). The color scale indicates the average Fitness between 1998 and 2018 (missing entries are colored in gray), with darker and lighter tone for higher and lower Fitness, respectively (the scale [0,1] is the same for the national and regional levels). Notice how the Fitness of a nation cannot be simply obtained by summing nor averaging the Fitness of its regions (see Fig. 3 below). The elliptic projection of the map follows the Robinson projection (esri:54030).

approach is also used in classic EFC literature, where Economic Fitness is scattered against a monetary measure of income, typically the Gross Domestic Product per capita. The dynamics in the two dimensional space (Fitness-Income) highlights clusters of similar Economies, allowing for a very precise Economic forecasting (Tacchella et al., 2018). Instead, in panel-(a) of Fig. 2 we employ the Higher Education expenditures on R&D (HERD) (OECD, 2020a), namely the expenditure for basic research performed in the higher education sectors, which, among the sources of public funding, are the most related with scientific performance as measured by citations of published documents.[7] This data is available only for OECD member countries and a few other important economies (such as China and Russia).Therefore, the following analysis focuses only on this subset of high and middle income countries. Notice that in previous work applying EFC to trade data, GDP is interpreted as a consequence of Fitness. In our analysis, HERD is more likely an input to Scientific Fitness. In spite of the difference in interpretation, we represent Fitness on the x-axis and HERD on the y-axis to allow easier visual comparison between this and other works in the EFC literature.

We observe that the most developed Economies tend to concentrate in the top right corner of the diagram in Fig. 2 (enlarged in the inset) characterized by high values of both Fitness and HERD-pc. Other nations are scattered along the diagonal, for which Scientific Fitness is proportional to resources invested, and their trajectories are typically directed towards the top-right region: these countries are quickly catching up with the most advanced ones. However, off-diagonal trajectories provide interesting information, similar to that obtained in EFC applications to trade data (Cristelli et al., 2013). The top left corner contains small national research

[7] The other sources of public funding are OECD, 2020a: Business Expenditures on R&D (BERD), namely R&D expenditures performed in the business sector, which is mostly related to the creation of new products and production techniques (patents); Government Intramural Expenditures on R&D (GOVERD), namely expenditures in the government sector, which is often mission-oriented and therefore less connected to publication outputs (see Cimini et al. (2016) and the discussion therein). We show in SI results of the analysis performed using Gross Expenditure on R&D (GERD), given by the sum of HERD, BERD and GOVERD.

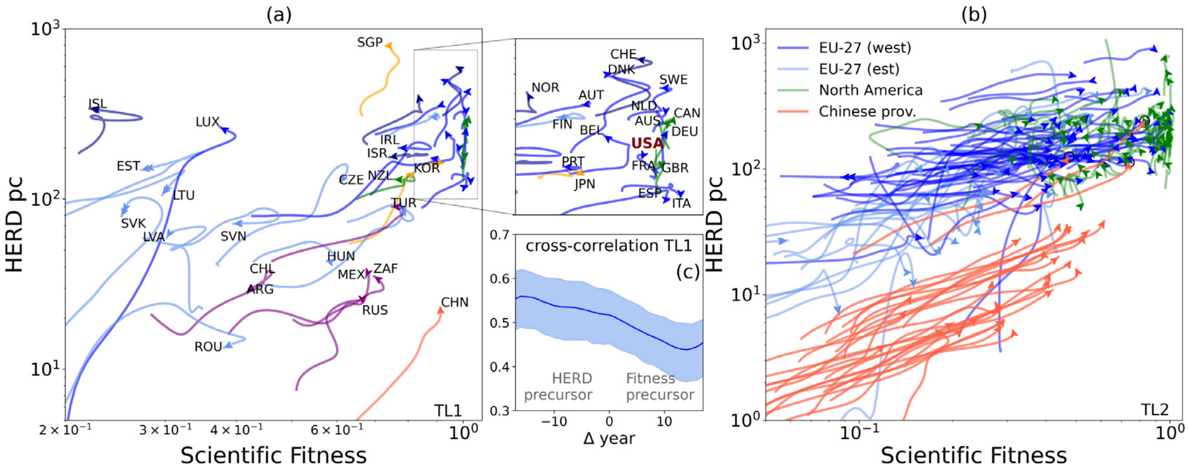

**Fig. 2.** (Panel a) Trajectories of nations (TL1) in the plane defined by Scientific Fitness and resources invested, the latter measured by Higher Education expenditures on R&D per capita (HERD-pc). Line colors are used to group nations into macro-areas: dark blue for west EU nations (plus Switzerland, Israel, Norway, Island), light blue for east EU nations, green for the English-speaking nations (United States, United Kingdom, Canada, Australia, New Zealand) red for China, yellow for the other Asian nations (Singapore, South Korea, Japan) and purple for middle-income countries (Russia, South Africa, Mexico, Argentina, Chile). Trajectories represent data from 2000 to 2017, with the arrow indicating the direction of time. The inset zooms on the top-right corner where there is a concentration of highly competitive nations. (Panel b) Trajectories are also displayed for regions (TL2) belonging to China and a selection of EU west, EU east and North America nations. (Panel c) Cross-correlation between Scientific Fitness and HERD at the national scale (TL1) averaged over the whole set of countries as a function of the temporal delay (Δ year) used to compute these quantities. The blue contour represents the 25 – 75% quantile, generated with a bootstrapping technique. Note that a cross-correlation value of about 0.5 is comparable to analogous estimations carried out in the economics context (Patelli et al., 2022).

systems with peculiar features, where investments are not efficiently turned into scientific competitiveness. This is for instance the case of Iceland, which does not attract much attention in terms of citations, and of Luxembourg, where the presence several private firms headquarters may bias the scientific production towards patent-related documents (Patelli et al., 2017). In the opposite corner, China (and to a lesser extent Russia, South Africa and Mexico) features a high scientific competitiveness despite low public R&D expenditures, with both quantities growing quickly in time.

### 4.2. Regional level analysis

A significant advantage and novelty of our framework is the possibility to perform the analysis of Scientific Fitness at a more detailed geographic level, in order to highlight the competitiveness of specific regions within nations. The bottom maps of Fig. 1 report the Scientific Fitness of regions, as defined by OECD Territorial Level 2 (TL2), for three macro-areas: North America, Europe and East Asia. We observe a recurrent pattern whereby the Fitness of a nation is mostly concentrated in its capital region (also because capitals usually host the headquarters of the largest national research institute), with the exception of the English-speaking nations, United States, United Kingdom, and Australia (not shown), featuring a rather uniform Fitness profile in their regions. Such a widespread competitiveness can be partially due to a language bias of the dataset, since English is predominant in Hard Science but less in Soft Science, especially for Social Sciences and Humanities (Sivertsen & Larsen, 2012). This bias might create an artificial advantage of native English speakers but it is not a dominant effect in the analysis, as shown in Methods and SI.

Focusing on the evolution of Scientific Fitness and HERD-pc at TL2, in the right panel of Fig. 2, we see again that most of the North American regions are top performers, while European regions form a cloud ranging from low to high competitiveness. Instead, China stands alone: only three provinces (Beijing, Hong Kong and Tianjin) belong to the cloud of EU regions, while the others follow a very regular flow with a steady increase parallel to the European cluster. Indeed, China invested enormously in science starting from the end nineteenth century, with growing R&D expenditure throughout the country. Apart from the three outliers, the competitiveness of Chinese provinces has not yet reached that of the western regions, but it will eventually (Huang, 2018; Xie and Freeman, 2019). This can be clearly seen in panel-(a) of Fig. 3, where the trajectories of regional Scientific Fitness are scattered with those of *document Fitness*, i.e., competitiveness computed on document production (see Methods). Mainland Chinese provinces follow a unique pattern, whereby their document Fitness has increased substantially from 2000 to 2018, due to growing resources and the consequent acquisition of new capabilities. However, at the beginning the research was not able to collect many citations, likely due to a low level of competitiveness. Only recently Chinese research became very competitive with a consequent growth in Scientific Fitness.

Furthermore, the analysis at TL2 indicates that the Fitness of a nation is not obtained by simply averaging the Fitness of its regions, because the most exclusive capabilities are typically concentrated only in a few regions. This is confirmed by panel-(b) of Fig. 3, which shows that the national Fitness is more correlated to the Fitness of its most competitive region (Pearson correlation 0.96) rather than

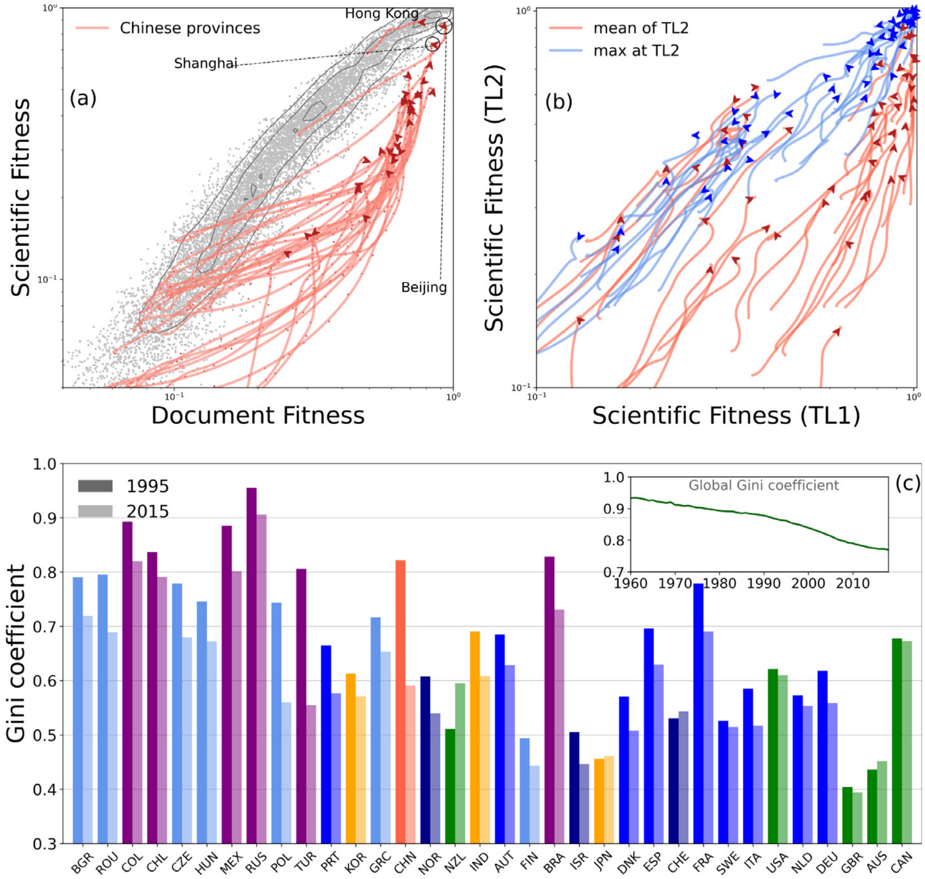

**Fig. 3.** (Panel a) Comparison of Scientific Fitness and document Fitness (i.e., Fitness computed using published documents) at the regional level (TL2). The black lines indicate the density level contour of the cloud of points while each red trajectory indicates the evolution of a Chinese province. The trajectories map the evolution from 2000 to 2018, with the arrow indicating the direction of time. (Panel b) Comparison between the Scientific Fitness of nations, computed either at the national level (TL1) or as the mean (red line) or maximum (blue line) of the Fitness of internal regions (TL2). (Panel c) Gini coefficients of each nation, computed over the citation counts of internal regions. We report values for two years: 1995 (full color bars) and 2015 (shade color bars). Nations are ordered according to their average Scientific Fitness in the central decade (2000–2010). The inset represents the temporal evolution of the Gini coefficient of the whole world.

to the mean Fitness of its regions (Pearson correlation 0.70). More importantly, our framework highlights the strong heterogeneity of Fitness values both across and within nations, consistently locating the geographic inequalities of the scientific research system. Values of the Gini coefficient in 1995 and 2005 (see Methods for the precise definition of the coefficients implemented) are shown in panel-(c) of Fig. 3 for the nations having more than 4 TL2 regions. The analysis highlights in general that English-speaking nations, e.g. UK and Australia, feature low inequalities, while mid-income countries are characterized by the highest inequality levels. We also compute the global Gini coefficient over all available regions in the inset of panel-(c) showing that the global level of inequality is slowly and steadily decreasing in time.

### *4.3. Scientific sector fitness*

Similarly to what we do with the geographic down-scaling, we can increase the resolution regarding the research sectors by exploiting the hierarchical classification of FoS. Thus, for example, instead of computing the total Scientific Fitness of a geographic area we can compute its *sector* Fitness (Operti et al., 2018; Barbieri et al., 2022) restricted to one of the 19 entries of the FoS main hierarchical level. Fig. 4 shows the radar plots of the sector Fitness for some example regions. The 19 research sectors are ordered clockwise in the radar according to their Complexity (see Methods), indicating that *Business* is the most Complex while *Material Science* is the less Complex FoS. Note that the EFC algorithm typically assigns higher Complexity to Soft Sciences (e.g. Social Sciences) rather than to Medical and Hard Sciences, because it turns out that only the most competitive players are active in the former sectors, while the latter sectors are more ubiquitous. In principle, this pattern can be caused by the aforementioned English bias in Soft Sciences, although a more fundamental explanation exists: only the most developed research systems have reached the capabilities required to perform researches in, e.g., *Business Administration, Environmental Ethics* and *Cognitive Science*. These sectors require solid

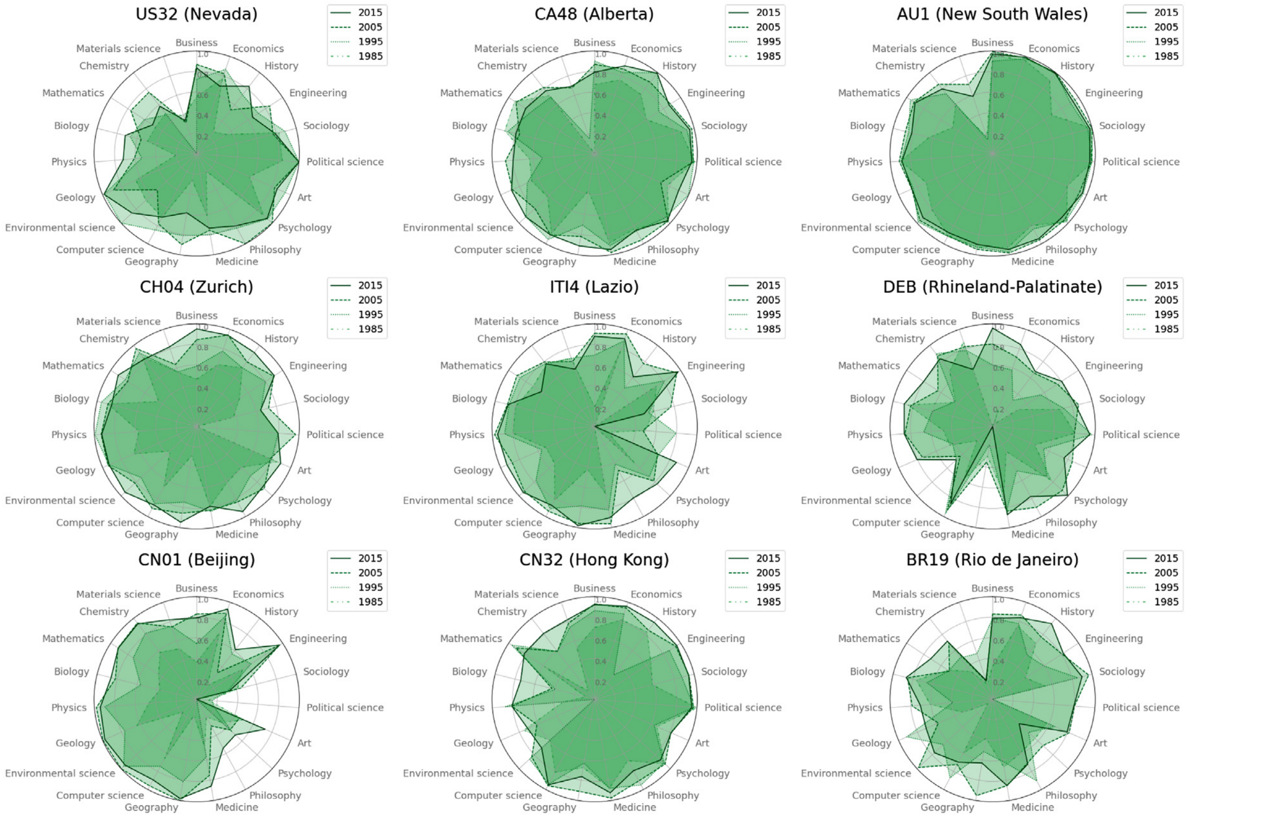

**Fig. 4.** Radar plots of the scientific sector Fitness of different sample regions (TL2). Top row: Nevada (USA), Alberta (Canada), New South Wales (Sydney). Central row: Zurich (Switzerland), Lazio (Italy), Rhineland-Palatinate (Germany). Bottom row: Beijing (China), Hong Kong (China), Rio de Janeiro (Brazil). Sectors are ordered clockwise with decreasing average complexity (Business is the most complex and Material Science is the less complex sector). The radar lines indicate how Fitness has evolved over the course of thirty years, from 1985 to 2015.

prerequisites in Hard Sciences as well as substantial available resources, and they are aimed at addressing the most advanced needs of the society (Cimini et al., 2014). Moreover, note that the average Complexity of a research sector does not fully reflect the Complexity of the associated sub-sectors, since also in Hard Sciences there are highly sophisticated sectors that require expensive instruments. For example, while the average Complexity of *Business* is about 1.82, the Complexity of *Polymer Science*, a child code of *Material Science* and *Chemistry*, is as high as 5.34. Overall, the analysis of the scientific sector Fitness allows to quantitatively detect the strengths and weaknesses of each region. For instance, Fig. 4 shows how Beijing experienced a fast growth in competitiveness in Hard Sciences while it still falls back in artistic and cultural areas. On the contrary, regions like Zurich, Lazio and Alberta have a more uniform pattern of competitiveness, especially in recent decades. Remarkably, also top-performing regions like New South Wales have competitive gaps but only in the less Complex sectors.

## 5. Conclusions

This work aims to bring together two recent lines of research: *Science of Science* (Fortunato et al., 2018; Waltman, 2016; Zeng et al., 2017), which develops quantitative methods and assessment tools to study the evolution of Science itself, and *Economic Fitness and Complexity* (Hidalgo, 2021; Hidalgo and Hausmann, 2009; Tacchella et al., 2012), which aims at measuring the productive capabilities of economic systems. Indeed, our framework to assess competitiveness in scientific research builds on the theory of hidden capabilities and employs properly calibrated bibliometric indicators. The proposed methods allow for a consistent comparison between different geographic areas and research sectors at varying level of resolution. We further characterized the performance of scientific actors across the various research sectors, and showed that the evolution of research systems can be properly described using two dimensions, Scientific Fitness and R&D expenditure. In the plane defined by these variables, nations form clusters of similar research systems operating within countries that have reached comparable stages of development.

Similarly to other the classic applications in the EFC literature, this study shows that a high explanatory power is achieved when Fitness is coupled with the amount of resources available in the system. Typically, the EFC literature proxies resource endowments with Gross Domestic Product (GDP) (Tacchella et al., 2018); for our purposes, HERD is the more appropriate measure. However, there is a fundamental difference between the use of GDP and HERD. GDP is a measure of generated wealth, hence it reflects outcomes of the production process and it can be interpreted as a consequence of Economic Fitness. Instead, HERD measures the amount of public resources that are fed into the scientific system and thus is an input requisite for Scientific Fitness. Consequently, while both the trajectories of countries in the GDP-Fitness plane and in the Scientific Fitness-HERD allow to extract interesting patterns concerning the way in which nations cluster in the plane, there are also remarkable differences in their interpretation. A comprehensive analysis of the relation between Scientific Fitness and different measures of input and output of research systems represents a promising avenue for future research.

Our analysis shows that EFC allows us to extract valuable information about Scientific capabilities at both a national and subnational level. However, it should be noted that there is a limit to the level of detail we can achieve. In fact, beyond a given threshold level of disaggregation in the definition of regions or scientific fields, the data eventually becomes too sparse to allow extracting a meaningful signal. A further limitation of the EFC approach is that the method naturally produces a ranking of fitness and complexity values, but is incapable of accounting for the global growth of the system. Therefore, the results yielded by the EFC approach cannot replace conventional analytical methods, but rather complement them by offering an alternative perspective. Finally, the EFC algorithm treats geographic areas as separate entities and infers the capabilities of each one based on how they diversify their output across fields of science. This implies that EFC potentially underestimates synergies and collaborations between geographic areas that are not encoded in the Scientific Bipartite Network linking areas to fields of science. Employing fractional counting in constructing the data to give all geographic areas credit for scientific publication by employing fractional counting helps mitigate this concern.

In addition to uncovering non-trivial patterns in the evolution of national and regional knowledge production systems, the application of the EFC methodology to the realm of scientific production data also has the potential relevance for policy making. Even though the direct concern of economic policy is not so much knowledge creation, but rather economic output or innovation, it is known that competitiveness in scientific fields is robustly linked to the development of competitive advantages in patenting as well as export (Pugliese et al., 2017). Since success in one of the above three layers – knowledge, innovation, trade – tends to be a precursor of success in the others, it is reasonable to argue that a long-sighted approach to growth and development policies can only benefit from factoring knowledge production capabilities into the equation. Finally, the analysis of the scientific competitiveness of regional areas add a tool in the analysis of local capabilities, necessary in the developments of less wealthy regions.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The views expressed in this article do not reflect those of the European Commission.

The datasets generated and analysed during the current study are available in the *Scientific database* repository, https://efcdata.cref.it/.

## CRediT authorship contribution statement

**Aurelio Patelli:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft. **Lorenzo Napolitano:** Conceptualization, Methodology, Data curation, Writing – original draft. **Giulio Cimini:** Conceptualization, Methodology, Formal analysis, Writing – original draft. **Andrea Gabrielli:** Conceptualization, Methodology, Writing – original draft, Supervision.

## Acknowledgments

A.P. gratefully acknowledge funding received from the Joint Research Centre of Seville (grant number 938036-2019 IT).

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2022.101357.

## References

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics, 101*(2), 1129–1144. 10.1007/s11192-014-1269-8.

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *J. Inform., 6*, 36–43.

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

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour, 4*(3), 248–254. 10.1038/s41562-019-0803-3.

Barbieri, N., Consoli, D., Napolitano, L., Perruchas, F., Pugliese, E., & Sbardella, A. (2022). Regional technological capabilities and green opportunities in europe. *The Journal of Technology Transfer*. 10.1007/s10961-022-09952-y.

Burrel, Q. L. (2002). The nth-citation distribution and obsolescence. *Scientometrics, 53*.

Cimini, C., Gabrielli, A., & Sylos Labini, F. (2014). The scientific competitiveness of nations. *PloS one, 9*, 12.

Cimini, G., Zaccaria, A., & Gabrielli, A. (2016). Investigating the interplay between fundamentals of national research systems: Performance, investments and international collaborations. *Journal of Informetrics, 10*, 200–211.

Cozza, C., & Schettino, F. (2015). In C. Mussida, & F. Pastore (Eds.), *Explaining the patenting propensity: A regional analysis using epo-oecd data* (pp. 219–236)). Springer Berlin Heidelberg.

Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G., & Pietronero, L. (2013). Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one, 8*.

Cristelli, M., Tacchella, A., & Pietronero, L. (2015). The heterogeneous dynamics of economic complexity. *PloS one, 10*, 2.

Dosi, G., Nelson, R., & Winter, S. (2000). *The nature and dynamics of organizational capabilities*. Oxford University Press.

Eom, Y., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PloS one, 6*.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. *Journal of Informetrics, 9*(4), 895–906. 10.1016/j.joi.2015.09.005.

Färber, M. (2019). The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. *ISWC', 19*, 113–129.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., … Barabasi, Albert-Laszlo, et al., (2018). Science of science. *Science (New York, N.Y.), 359*(6379). 10.1126/science.aao0185.

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics, 3*(2), 92–113. 10.1038/s42254-020-00275-1.

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences of the United States of America, 106*.

Huang, F. (2018). Quality deficit belies the hype. *Nature, 564*, S70–S71.

Hug, S. E., & Brändle, M. P. (2017). The coverage of microsoft academic: Analyzing the publication output of a university. *Scientometrics, 113*(3), 1551–1571. 10.1007/s11192-017-2535-3.

Iacopini, I., Di Bona, G., Ubaldi, E., Loreto, V., & Latora, V. (2020). Interacting discovery processes on complex networks. *Physical Review Letters, 125*, 248301. 10.1103/PhysRevLett.125.248301.

King, D. (2004). The scientific impact of nations. *Nature, 430*, 311–316. 10.1038/430311a.

Laverde-Rojas, H., & Correa, J. C. (2019). Can scientific productivity impact the economic complexity of countries? *Scientometrics, 120*(1), 267–282.

Mariani, M., Ren, Z.-M., Bascompte, J., & Tessone, C. J. (2019). Nestedness in complex networks: Observation, emergence, and implications. *Physics Reports, 813*.

May, R. M. (1997). The scientific wealth of nations. *Science (New York, N.Y.), 275*(5301), 793–796. 10.1126/science.275.5301.793.

Medo, M., & Cimini, G. (2016). Model-based evaluation of scientific impact indicators. *Phys. Rev. E, 94*, 032312. 10.1103/PhysRevE.94.032312.

Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters, 107*, 238701. 10.1103/PhysRevLett.107.238701.

Operti, Felipe G., Pugliese, Emanuele, Andrade, Jose S. Jr., Pietronero, Luciano, Gabrielli, Andrea, et al., (2018). Dynamics in the Fitness-Income plane: Brazilian states vs World countries. *PloS one, 13*(6). 10.1371/journal.pone.0197616.

OECD (2020a). Gross domestic spending on R&D (indicator). doi:10.1787/d8b068b4-en.

OECD (2020b). OECD Territorial grids. *Technical document by the OECD Centre for Entrepreneurship, SMEs, Regions and Cities*. Available online at: http://www.oecd.org/regional/regional-statistics/territorial-grid.pdf.

Patelli, A., Napolitano, L., Zaccaria, A., Cimini, G., Gabrielli, A., & Pietronero, L. (2021). Jrc final report: *Economic Fitness and Complexity: an inquiry into the innovation and competitiveness of world regions*.

Patelli, A., Cimini, G., Pugliese, E., & Gabrielli, A. (2017). The scientific impact of nations on scientific and technological development. *Journal of Informetrics, 11*, 1229–1237.

Patelli, A., Zaccaria, A., & Pietronero, L. (2022). Integrated database for economic complexity. *Scientific Data, 9*, 628. 10.1038/s41597-022-01732-5.

Petersen, Alexander M., Pan, Raj K., Pammolli, Fabio, Fortunato, Santo, et al., (2019). Methods to account for citation inflation in research evaluation. *Research Policy, 48*(7), 1855–1865. 10.1016/j.respol.2019.04.009.

Pugliese, E., Cimini, G., Patelli, A., Zaccaria, A., Pietronero, L., & Gabrielli, A. (2017). Unfolding the innovation system for the development of countries: Co-evolution of science, technology and production. *Scientific Reports, 9*.

Pugliese, E., Zaccaria, A., & Pietronero, L. (2016). On the convergence of the fitness-complexity algorithm,. *The European Physical Journal. Special Topics, 225*, 1893.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America, 105*.

Romeo, M., Da Costa, V., & Bardou, F. (2003). Broad distribution effects in sums of lognormal random variables. *European Physical Journal B: Condensed Matter and Complex Systems, 32*, 513–525.

ScimagoLab (2020). https://www.scimagojr.com/countryrank.php.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Sivertsen, G., & Larsen, L. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*, 567–575.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2012). A new metrics for countries' fitness and products complexity. *Scientific Reports, 2*.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2013). Economic complexity: Conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control, 37*.

Tacchella, A., Mazzilli, D., & Pietronero, L. (2018). A dynamical systems approach to gross domestic product forecasting. *Nature Physics, 14*, 861–865.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the fourteenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 990–998).

Tria, F., Loreto, V., Servedio, V., & Strogatz, S. H. (2014). The dynamics of correlated novelties. *Scientific Reports, 4*. 10.1038/srep05890.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*, 365–391.

Wang, D., Song, C., & A-L, B. (2013). Quantifying long-term scientific impact. *Science (New York, N.Y.), 342*, 127–132.

Xie, Q., & Freeman, R. B. (2019). Bigger than you thought: China's contribution to scientific publications and its impact on the global economy. *China and World Economy, 27*.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports, 714–715*, 1–73. 10.1016/j.physrep.2017.10.001. The Science of Science: From the Perspective of Complex Systems