Measuring and interpreting the differences of the nations' scientific specialization indexes by output and by input

Giovanni Abramo[1] iD, Ciriaco Andrea D'Angelo[1,2] iD, and Flavia Di Costa[2]

[1]Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Rome, Italy
[2]Department of Engineering and Management, University of Rome "Tor Vergata," Rome, Italy

ABSTRACT

This paper compares the national scientific profiles of 199 countries in 254 fields, tracked by two indices of scientific specialization based respectively on indicators of input and output. For each country, the indicator of inputs considers the number of researchers in each field. The output indicator, named *Total Fractional Impact*, based on the citations of publications indexed in the Web of Science, measures the scholarly impact of knowledge produced in each field. For each country, the approach allows us to measure the deviations between the two profiles, thereby revealing potential differences in research efficiency and/or capital allocation across fields, compared to benchmark countries.

1. INTRODUCTION

Policy-makers who have knowledge of the scientific specializations of their country can better formulate research policies and funding priorities, including by specific field, and can better assess the effectiveness of their initiatives in relation to strategic priorities. Whether public or private, however, stakeholders face major challenges in identifying scientific priorities and then parceling their investments (King, 2004; May, 1997). What is necessary is not only knowledge of the home nation scientific profile but also its relation to those of other countries, at regional and global levels.

The measurement of research activity and the construction of a national scientific profile can be carried out by considering either the input employed (resources and capital investment, research personnel, etc.) or the output produced (know-how, scientific publications, patents, etc.); that is, the knowledge developed and its scholarly impact (Sugimoto & Larivière, 2018).

In a previous work, for purposes of tracing the scientific profiles of countries, we proposed an index of scientific specialization based on scholarly impact of 2010–2019 Web of Science (WoS) publications in each subject category (SC) (Abramo, D'Angelo, & Di Costa, 2022a). By producing a specialization profile for each country in relation to all SCs (254), we were able to identify the distinctive characteristics of individual countries and country clusters.

However, if we consider the whole process of scientific research production as a black box, the calculation of specialization indices can also be carried out by considering input indicators

alongside the output indicators. The former approach traces the profile of a country through the sectoral distribution of research investments; the latter through the relative distribution of its scientific production.

From an operational point of view, tracing the research profile of a country on the basis of input indicators is a challenging task, because at the global level, gathering input data disaggregated by field is formidable, even more so by univocal classification of those fields. Input data, or production factors according to the microeconomic theory of production, are labor (L) and capital (K); that is, all resources other than labor used to conduct research activities. While K data are not available, in this paper we go some way to overcoming the obstacle concerning L data. In fact, the bibliometric approach allows not only measurement and classification of output, through observation of scientific output, but indirectly also the input, limited to the research staff. In fact, having understood how to disambiguate authors' identities and their country affiliations, this makes it possible to measure the size of the research staff of a country and to classify it per SC based on the prevalent SC in which each author's publications fall. It is then possible to measure the scientific specialization of countries with input data (limited to L), in a similar way as with output data.

It is then interesting to check whether and to what extent the resulting scientific profiles are different. The share of research fields showing deviations between the two indices would reveal differences in research efficiency and/or allocation of K across fields, compared to benchmark countries. In fact, because research output is a function of L and K, if a field specialization index is higher by input than by output, a possible explanation is that the country has historically invested less K in that field than in others and/or that the productivity of the researchers, compared to other countries, is lower in that field. When the share of such fields surpasses one half, the inference would be that the country is entering the area of imbalance across fields, in the efficiency of their research and/or capital allocations. Were K data available and accounted for, those differences would reveal directly field-level comparative advantages across countries.

Essentially, to move the national research profile towards alignment with strategic objectives, governments can act on two levers: differentiated allocation of public funds across fields, and/or differentiation of productivity incentives by scientific fields, although the latter would not be easy in practice. In any case, the effects of these interventions on field outputs of research, and on shifting the scientific profile, is in part dependent on the status of productivity across these very fields.

The objectives of the present work are therefore, for each country

- produce two specialization profiles, respectively based on input and output indicators, corresponding to each of the 254 SCs of the WoS classification scheme;
- analyze the two specialization profiles of countries by input and output indicators; and
- assess the deviations between the two profiles for individual countries and country clusters;

all this in a manner supportive of policy-makers intending to formulate research policies and priorities for funding by field.

The next section of this paper reviews the relevant literature. Section 3 describes the data and indicators used for analysis, and the methodology adopted for construction of the specialization profiles. Section 4 presents the results of the analysis and Section 5 comments the main findings and discusses the policy implications.

2. LITERATURE REVIEW

Scholars have generally applied frameworks from business or economics in studying specialization levels in scientific research. The most common approach is by "revealed comparative advantage" (Aksnes, Sivertsen et al., 2017; Allik, Realo, & Lauk, 2020; Bongioanni, Daraio et al., 2015; Cimini, Zaccaria, & Gabrielli, 2016; Horta & Veloso, 2007; Leydesdorff & Wagner, 2009; Li, 2017; Patelli, Cimini et al., 2017; Sandström & Van den Besselaar, 2018). Examining a field at international level, this approach "reveals" the comparative advantage of a country in proportions of labor factor, or output produced, compared globally or to a selection of countries. All comparative advantage indices used in international economics originate from the Balassa or "RCA" index (Balassa, 1965). The first to transfer RCA to investigation of specialization in scientific research was Frame (1977), who introduced the so-called "activity index.[1]" This indicator is typically based on one of the easily measured macroscopic bibliometric variables: total publications from a country; total citations received by the country's publications (Aksnes, van Leeuwen, & Sivertsen, 2014; Harzing & Giroud, 2014); and in some case more sophisticated combinations of output and impact (Abramo, D'Angelo, & Di Costa, 2014; Abramo et al., 2022a).

The value of the activity index is given by the ratio of two ratios. The first one measures the share of research effort (or output) of a country in a given field with respect to the national total, and the second one measures the same share but at a global level. The indicator is expressed as an absolute value or transformed on a scale [−100; +100] for easier understanding and comparison.

Subsequent to detailed analysis of its technicalities, Glänzel (2000), and Schubert and Braun (1986) have provided interpretations of this indicator. Other authors have explored theoretical problems in the construction of the activity index and related indicators (Aksnes et al., 2014; Rousseau, 2018, 2019; Rousseau & Yang, 2012).

The bibliometric indicators generally used are based on output data extracted from bibliographic repertories (WoS, Scopus) which, despite coverage problems (by discipline, language, country, etc.), have become the de facto standard for measuring research, and more generally, for studies in the field of the so-called "science of science" (Archambault, Vignola-Gagné et al., 2006; Hicks, 1999; Waltman, 2016). Compared to other approaches of measuring research, bibliometrics clearly has the advantage of access to data, gathered by repository publishers according to globally standardized procedures.

In contrast, input data are generally collected through local and international surveys, under the auspices of national research councils or international organizations, such as OECD and UNESCO. Although such entities collect and regularly update their data, none have the mandate or capacities to apply standard classification systems, so none can provide data sufficient for reliable study of specialization. Given the inaccessibility of data on inputs, scholars interested in the investigation of specialization at macro (i.e., country) level have thus far engaged solely with data on outputs.

On the other hand, there is no shortage of analyses on input and output data at meso level (i.e., surveys of data on a small set of local institutions, enabling evaluation of their specialization). Heinze, Tunger et al. (2019), for example, described research and teaching profiles for 68 public universities in Germany (from 1992 to 2015) and produced specialization maps for

---

[1] Activity index (AI) was originally defined as the ratio between the country's share in the world's publication output in the given field and the country's share in the world's publication output in all fields (Frame, 1977).

each of them. Fuchs and Heinze (2021) then revised the analysis on an updated data set (1992 to 2018). Teixeira, Rocha et al. (2012) adapted one output and three input measures from the RCA index of Balassa (1965) in the study of field-by-field diversity (specialization and/or diversification) of Portuguese higher education institutions.

Thus far however, in measurement of specialization at macro/country level, for the reasons explained above, there remain no works using input data. In this paper we try to fill this gap, using the bibliometric approach.

3. DATA AND METHODS

Observing the authorship of scientific publications, then taking on the task of disambiguating the author identities, and tagging by country affiliation and field of specialization, we are ultimately able to measure the size of a country's research staff in a given field. This input measure can then be used to construct the country's sectoral specialization profile in terms of inputs, in the manner of traditional approaches dealing only with outputs. In the following, we explain the methodological details.

The data set for the analysis is the same as previously used by Abramo et al. (2022a), which applied the rule-based scoring and clustering algorithm of Caron and van Eck (2014) to data extracted from the in-house WoS database of the Centre for Science and Technology Studies (CWTS) at Leiden University (updated to the 13th week of 2021). For this algorithm, bibliometric metadata on authors and their publications are taken as input, and clusters of publications likely to be written by the same author are taken as output. The algorithm considers four categories of bibliographic elements:

- author name (first and last name, affiliation, email);
- article (shared coauthors, grant numbers, address not linked to authors);
- source (SC, journal); and
- citation (self-citations, cocitations, bibliographic coupling).

The higher the number of shared bibliographic elements (source, topic, coauthors, emails, affiliations, references, etc.) between two publications, the stronger is the evidence that these are written by the same author.

Based on scoring values and thresholds, defined on a verified seed set, the algorithm develops clusters of publications and assigns them to an individual.

Of course, the algorithm is far from being error free, especially for authors with popular names, or production of highly diversified and heterogeneous bibliographic elements, a circumstance that could lead to splitting their portfolio in two or more clusters.

However, at the aggregate country level, this latter error, as extensively explained in the theory and methodology of the previous work, will have only marginal effects on analytical results. Referring to Abramo et al. (2022a), an important note is that to increase robustness of the analysis, the data set excludes those clusters that fail to comply with one or more of the following conditions:

- contain at least 10 publications (excludes "occasional" researchers, for whom clustering has lower confidence levels);
- of which at least one publication is after 2018 (designed to exclude researchers no longer active); and

- with a "research age"[2] of minimum 5 years (designed to include only "established" researchers).

Through such "cuts" we effectively exclude small clusters, related to very young or occasional researchers but also those related to researchers no longer active (e.g., who are now retired). We also exclude part of those clusters deriving from the splitting of authors with popular names and/or with highly diversified scientific production, caused by the Caron and van Eck algorithm. All this allows us to have a higher confidence that the resulting data set actually represents the research staff of a given country, at present.

The final data set consists of over 2 million clusters, accounting for over 120 million authorships, related to almost 17 million unique publications. On average each cluster contains 58 publications, and each unique publication is coauthored by eight distinct clusters.

For field classification purposes, we use the WoS scheme, including 254 SCs[3]. Each cluster in the data set is provided with the 2010–2019 related WoS indexed publications[4] and is associated with a field, given by the "prevalent" SC of its publications (i.e., the one hosting most of his or her scientific production)[5]. In the input-based approach, the specialization index $(IB)SI_{jk}$ of country k, in the SC j is

$$(IB)SI_{jk} = \frac{RS_{jk}}{\sum_j RS_{jk}} \Big/ \frac{\sum_k RS_{jk}}{\sum_j \sum_k RS_{jk}}, \tag{1}$$

where RS_{jk} = research staff, operationalized as number of clusters of the country k in the SC j.

The higher the value of SI_{jk} compared to 1, the more specialized the country k is in SC j, as the share of its research staff is higher than the expected value observed at world level. If SI_{jk} is less than 1 it means that no specialization is involved in SC j for country k.

In the output-based approach, instead, we use the composite indicator proposed in Abramo et al. (2022a), and called Total Fractional Impact (*TFI*), which is a combination of publication volume and field normalized citation impact. The *TFI* of a country k in SC j, is defined as

$$TFI_{jk} = \sum_{i=1}^{N_{jk}} f_{ik} \frac{c_i}{\bar{c}_j}, \tag{2}$$

where

N_{jk} = number of publications of country k, in SC j

f_{ik} = fractional contribution of coauthors of country k to publication i. For a publication with n coauthors, m of which are affiliated to country k, f_{ik} is equal to m/n[6]

c_i = citations received by publication i (counted at the 13th week of 2021)

$\bar{c}_j$ = average citations received by all cited publications of the same year and SC j of publication i[7]

---

[2] Given by the difference between the first and the last publication year assigned to the cluster.

[3] In WoS each publication inherits the *SC* of the hosting journal.

[4] Only articles, reviews, letters, and proceedings papers.

[5] Clusters with more than one prevalent SC are around 2% and are counted multiple times.

[6] Note that according to the CvE algorithm, each cluster (and thus each author) is associated with one and only one country.

[7] Abramo, Cicero, and D'Angelo (2012) demonstrated that the average of the distribution of citations received for all cited publications of the same year and SC is the best-performing scaling factor.

Applying Total Fractional Index, we can measure the output-based index of specialization $(OB)SI_{jk}$ of country k in SC j as

$$(OB)SI_{jk} = \frac{TFI_{jk}}{\sum_j TFI_{jk}} \Big/ \frac{\sum_k TFI_{jk}}{\sum_j \sum_k TFI_{jk}}. \tag{3}$$

In this case a value higher than 1 implies that country k is specialized in SC j, as the share of TFI in such SC is higher than the expected value observed at world level, and vice versa.

Countries can be more or less concentrated (diversified) in terms of scope (number of SCs) of research. We will assess that by the Gini index, or Gini coefficient, which measures variable distribution across a population (Gini, 1921). A higher Gini coefficient indicates greater inequality in the distribution of input (output) across SCs, with high-input SCs receiving much larger shares of the total input for research. The Gini coefficient ranges from 0 to 1, with 1 representing perfect inequality (concentration) and 0 representing perfect equality (diversification).

4. RESULTS

The analyses of the current paper, as follows, are aimed at comparing the distributions of SI_{jk} calculated from input and output data. For this, we construct 199 × 254 matrices containing the SI values, by input and output, for a set of 199 countries in each of the 254 WoS SCs. For reasons of space, we present only a few examples of possible data elaborations. The complete data on all 199 countries in 254 SCs are found in Abramo, D'Angelo, and Di Costa (2022b).

As a first example, Figure 1 shows, for China, the distribution of SIs detected for the SCs of Biomedical Research (14 in all). The SI values measured through output are never greater than unity; instead, when measured through input, five fields out of the 14 reach levels greater than unity. The $(OB)SI$ values are higher than the $(IB)SI$ values in only four cases: Among these, the highest absolute values are in Toxicology (0.759 by output data, 0.639 by input data). In absolute value, the greatest gap is in Medical Laboratory Technology (0.882 vs. 2.859), followed by Virology (1.136 vs. 0.592) and Oncology (1.175 vs. 0.678). It therefore emerges that for China, in general, there is a significant lack of specialization in this set of SCs, and above all a gap in capital investment and/or productivity, compared to other countries.

Figure 2 shows the comparison for the United States, looking at the SI values for input and output in the 20 SCs that are greatest by world output. In 15 out of 20, the $(OB)SI$ value is higher than the $(IB)SI$ value based on input, with a maximum deviation in Medicine, General & Internal; in this field, for the United States, the $(OB)SI$ is 1.368, compared to an SI by input of 0.831. At the opposite extreme for these 20 SCs is Chemistry, Multidisciplinary which shows an $(IB)SI$ of 1.267 versus an $(OB)SI$ of 0.743 by output, or in other words, 41% less. Also for the United States, whether for specialization index by input or output, there are nine SCs with values greater than unity, and of these, eight SCs represent the particular case where both SI values are greater than unity (Astronomy & Astrophysics; Biochemistry & Molecular Biology; Cardiac & Cardiovascular Systems; Clinical Neurology; Neurosciences; Oncology; Public, Environmental & Occupational Health; Surgery). For these eight SCs, the percentage variation between the two SI values was within the ±10% in 10 out of 20 cases.

Figure 3 instead examines Biochemistry & Molecular Biology, looking at the 20 largest countries by overall world share in the SC. For these, the radar graph shows a mismatch in the values of the specialization indices for some countries: especially for Russia (1.865 by
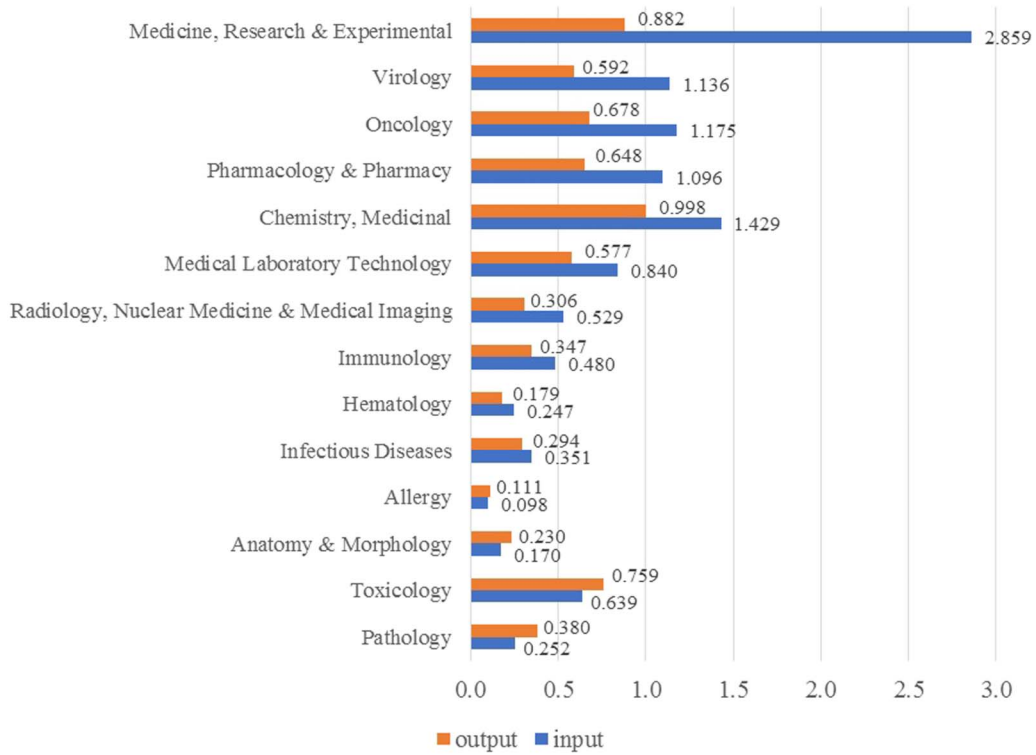
Figure 1. China: specialization indices for the subject categories in "Biomedical research".
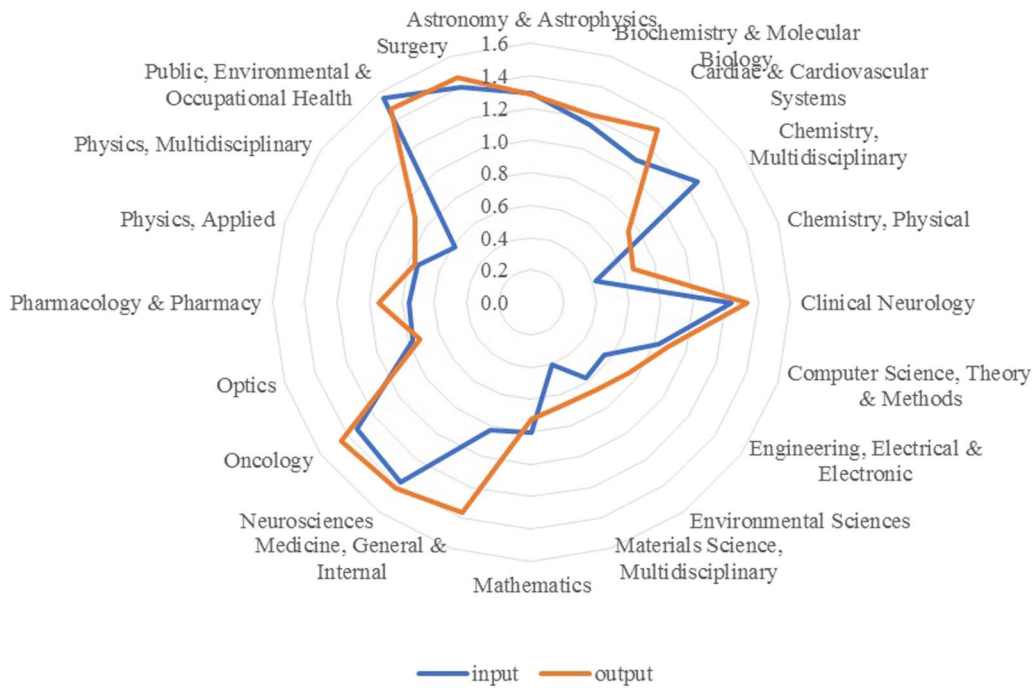

Figure 2. United States: specialization indices for the 20 subject categories that are largest by world output.
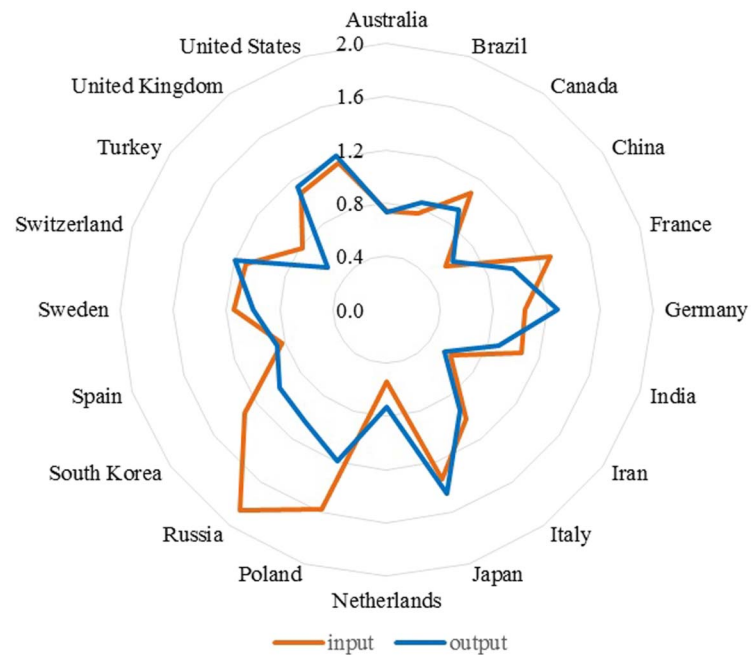
Figure 3. Biochemistry and Molecular Biology: specialization indices of the 20 largest countries by world share of output.

input vs. 1.040 by output), followed by Poland (1.572 vs. 1.191) and South Korea (1.316 vs. 0.911). Eight other countries on the list have *SI* values by input that are higher than those calculated by output; the opposite relation is seen in nine countries. The difference between values of the indicator falls in ±10% for eight countries out of the total 20 (Australia, Iran, Italy, Japan, Spain, Switzerland, United Kingdom, United States).

Table 1 provides an examination of the specialization profiles for the major European countries in terms of research output, specifically their top five SCs by specialization index based on input ((*IB*)*SI*) and output data ((*OB*)*SI*). All five of these European countries show a strong presence of "top" SCs (about 1/3 of the total, for both input and output) in the humanities and social sciences. Also interesting is that the intersection between the two sets of categories is rather limited: For France, Germany and Netherlands, two SCs appear in both columns; Italy and Spain have only one with a double appearance, and the United Kingdom has none. Finally, in this table, the top values of (*IB*)*SI* are greater than the corresponding top values of (*OB*)*SI* in 24 of the 30 total cases.

In Table 2, for the top seven countries by share of output, we look into the two SCs characterized by maximal difference between (*IB*)*SI* and (*OB*)*SI*, both negative and positive. In other words, for each country, columns 2–3 report the SCs with evident gaps in either or both of capital investment and productivity, given that the specialization indexes by output data do not align with what emerges concerning inputs. For China, for example, the maximal negative case ((*OB*)*SI* − (*IB*)*SI*) is found in Medicine, Research & Experimental, and in Mathematics, Interdisciplinary Applications; for Russia, this is found in Chemistry, Applied and Mining & Mineral Processing.

Columns 4–5 report the opposite situation (i.e., SCs with maximal difference of *SI* by output data over input data), evidently due to higher capital allocation and/or productive efficiency

Table 1.   Major European countries: top five SCs by specialization indices

Country	Input data SC	SI_{jk}	Output data SC	SI_{jk}
France	Acoustics	2.997	Literary Reviews	2.584
	Imaging Science & Photographic Technology	2.700	Critical Care Medicine	2.315
	Critical Care Medicine	2.369	Logic	2.271
	Mechanics	2.299	Geochemistry & Geophysics	2.230
	Geochemistry & Geophysics	2.031	Physics, Fluids & Plasmas	2.079
Germany	Literature, German, Dutch, Scandinavian	8.230	Literature, German, Dutch, Scandinavian	8.116
	Medical Ethics	7.091	Psychology, Psychoanalysis	3.331
	Psychology, Psychoanalysis	3.190	Microscopy	2.441
	Social Sciences, Mathematical Methods	3.124	Radiology, Nuclear Medicine & Medical Imaging	1.982
	Psychology, Educational	2.793	Dermatology	1.968
Italy	Instruments & Instrumentation	3.124	Art	3.180
	Geography, Physical	3.035	Architecture	3.170
	Architecture	2.810	Andrology	2.716
	Mineralogy	2.790	Medical Laboratory Technology	2.239
	Limnology	2.598	Engineering, Geological	2.212
Netherlands	Development Studies	6.191	Psychology, Mathematical	4.365
	Psychology, Mathematical	6.170	Public Administration	4.123
	Ethnic Studies	5.060	Regional & Urban Planning	3.573
	Social Sciences, Mathematical Methods	4.793	Primary Health Care	3.523
	Public Administration	4.198	Social Issues	3.189
Spain	Literary Theory & Criticism	7.705	Literature, Romance	10.502
	Psychology, Biological	4.220	Food Science & Technology	3.095
	Literature, Romance	3.566	Horticulture	2.501
	Psychology, Multidisciplinary	3.554	Agriculture, Multidisciplinary	2.436
	Education & Educational Research	3.412	Ornithology	2.269
United Kingdom	Ethnic Studies	7.587	Dance	7.250
	Development Studies	6.295	Literature, British Isles	6.601
	History of Social Sciences	5.966	Theater	6.184
	Social Sciences, Biomedical	5.861	Cultural Studies	5.531
	Classics	5.646	Medieval & Renaissance Studies	5.068

Table 2. Subject categories with min(max) $(OB)SI − (IB)SI$ difference for the top seven countries by share of output

Country	SC	$(OB)SI − (IB)SI$	SC	$(OB)SI − (IB)SI$
China	Medicine, Research & Experimental	−1.977	Computer Science, Cybernetics	+2.182
	Mathematics, Interdisciplinary Applications	−1.860	Physics, Condensed Matter	+1.186
France	Literature, British Isles	−1.378	Logic	+2.271
	Imaging Science & Photographic Technology	−1.255	Literature	+1.025
Germany	Medical Ethics	−6.423	Psychology, Biological	+1.004
	Social Sciences, Mathematical Methods	−2.131	Quantum Science & Technology	+0.977
Japan	Engineering, Ocean	−1.605	Cell & Tissue Engineering	+1.253
	Limnology	−1.524	Quantum Science & Technology	+1.240
Russia	Chemistry, Applied	−13.760	Literature, Slavic	+11.708
	Mining & Mineral Processing	−4.691	Paleontology	+2.255
United Kingdon	Ethnic Studies	−4.696	Poetry	+3.173
	Social Sciences, Biomedical	−3.561	Medieval & Renaissance Studies	+2.649
United States	Education, Special	−1.844	Limnology	+0.809
	Poetry	−1.518	Anatomy & Morphology	+0.620

compared to other SCs. For China, for example, such virtuous cases occur in Computer Science, Cybernetics and in Physics, Condensed Matter, while for the United States in Limnology and in Anatomy & Morphology.

Table 3 reports, for each of the top 20 countries by share of output, the shares of SCs with $(IB)SI$ greater than unity; $(OB)SI$ greater than unity; and $(OB)SI$ greater than $(IB)SI$. Within this group of 20 we quickly note some G7 countries, such as the United States, United Kingdom, Germany, and Canada, at the bottom of the table, but also another G7 country—Italy—near the top of the list. The first four countries in the list have about 70% of SCs with $(OB)SI$ greater than $(IB)SI$, the last four about 50%. It should be noted, however, that the latter case describes capital allocation and efficiency of research that are more balanced across fields.

4.1. Concentration/Diversification in Country Disciplinary Profiles

The disciplinary profile of a country can be more or less specialized in a few *SI*s or distributed in many (diversified or "balanced"). In this regard, there are interesting differences between countries when considering *SI*s based on input or output data. Table 4 shows, for the top 20 countries by share of output, the value of the GINI coefficient (output data) and the relative coefficients of variation of the distributions of SI values for the 254 SCs (input and output data). For all 20 countries except Iran, the GINI value for their $(IB)SI$ distribution is greater than the value for $(OB)SI$. Russia, Iran, and India, in view of the high values of GINI coefficients calculated in both modes, are the countries with highest level of concentration of sectoral

Table 3. Share of subject categories with (*IB*)*SI* and (*OB*)*SI* above one, and (*OB*)*SI* higher than (*IB*)*SI* for top 20 countries by share of output

Country	No of SCs*	Of which with (*IB*)*SI* > 1 (%)	Of which with (*OB*)*SI* > 1 (%)	Of which with (*OB*)*SI* > (*IB*)*SI* (%)
Turkey	219	33.8	42.5	83.1
Italy	238	35.7	42.9	76.5
Brazil	222	29.3	33.3	72.5
Poland	222	35.1	41.0	71.6
Russia	207	30.9	28.0	69.6
India	210	30.5	34.8	67.1
Japan	224	25.9	31.3	66.1
Switzerland	234	37.2	39.7	62.0
France	236	36.0	34.3	61.4
South Korea	219	32.0	33.3	61.2
Netherlands	246	46.7	55.3	61.0
Sweden	234	47.0	52.1	60.7
Iran	202	41.1	38.1	59.9
Spain	242	41.3	45.5	58.7
Germany	248	37.1	38.3	57.7
Australia	248	55.2	58.9	55.2
United Kingdom	250	54.4	57.6	50.8
United States	254	54.3	50.4	50.4
China	232	34.5	30.6	50.0
Canada	250	57.2	56.8	49.2

* With at least one researcher.

specializations. By contrast, the lowest values are recorded for the United States and Canada. Examining still further, Russia not only has the highest values of both GINI indicators (i.e., the profile strongest in specialization) but, along with China, India, and Iran, also has the lowest differences between the two values (ΔGINI 0.444). Basically, in all four of these countries, input and output are concentrated in certain fields functional to a specific industrialization model, most probably of historic character. The contrary situation of great difference between input and output distribution is observed for Switzerland (0.457 vs. 0.313), Sweden (0.412 vs. 0.279), and Turkey (0.579 vs. 0.466). On observing the variation coefficient, instead of GINI, similar trends in disciplinary profiles emerge: The largest differences between coefficients for distribution of (*IB*)*SI* and (*OB*)*SI* are for Switzerland and Poland; the smallest for Russia and Iran. Overall (Table 4) the values of variation coefficient fall in the intervals 0.604–2.025 for (*IB*)*SI* and 0.439–2.220 for (*OB*)*SI*.

Table 4. Dispersion of national disciplinary profiles and GINI concentration indexes for top 20 countries by share of output

Country	Input data		Output data			
	GINI coefficient	Variation coefficient	GINI coefficient	Variation coefficient	ΔGINI	Δ Variation coefficient
Russia	0.750	2.025	0.706	2.220	0.044	−0.195
Iran	0.599	1.248	0.607	1.262	−0.008	−0.014
India	0.595	1.182	0.576	1.137	0.019	0.045
Brazil	0.580	1.434	0.519	1.186	0.061	0.248
China	0.540	1.020	0.517	0.962	0.023	0.058
Poland	0.576	1.515	0.471	1.000	0.105	0.515
Japan	0.513	0.958	0.466	0.850	0.047	0.108
Turkey	0.579	1.197	0.466	0.971	0.113	0.226
South Korea	0.533	1.111	0.460	0.878	0.073	0.233
Netherlands	0.440	0.873	0.363	0.661	0.077	0.212
United Kingdom	0.416	0.865	0.351	0.749	0.065	0.116
France	0.395	0.709	0.324	0.583	0.071	0.126
Switzerland	0.457	1.552	0.313	0.591	0.144	0.961
Italy	0.408	0.756	0.311	0.569	0.097	0.187
Australia	0.394	0.821	0.300	0.564	0.094	0.257
Spain	0.366	0.791	0.291	0.751	0.075	0.040
Germany	0.372	0.872	0.289	0.684	0.083	0.188
Sweden	0.412	0.812	0.279	0.544	0.133	0.268
Canada	0.356	0.922	0.244	0.460	0.112	0.462
United States	0.327	0.604	0.243	0.439	0.084	0.165

Figures 4 and 5 compare the national disciplinary profiles of the United States and Russia, the two countries already noted at the antipodes in specialization/differentiation of scientific profiles in terms of (*IB*)*SI* and (*OB*)*SI*. A first observation is that for both indices, the values for the United States never exceed 4.5. On the contrary, the trends for Russia show pronounced oscillations: (*IB*)*SI*, while in the range 0–4 for 237 of the 254 SCs, presents a number of sharp peaks, two of which are close to the value 16; for (*OB*)*SI* the trend is to even more oscillations, although with peaks not surpassing 8.

Finally, we investigated the relationship between the dispersion of the national profiles of the top 20 countries by share of output and the balance of efficiency of research and/or capital allocation across fields. The correlation analyses showed that countries with high dispersion are those more balanced (for (*IB*)*SI*, Pearson correlation coefficient: 0.543; Spearman correlation coefficient: 0.583; for (*OB*)*SI*, 0.420 and 0.514, respectively).
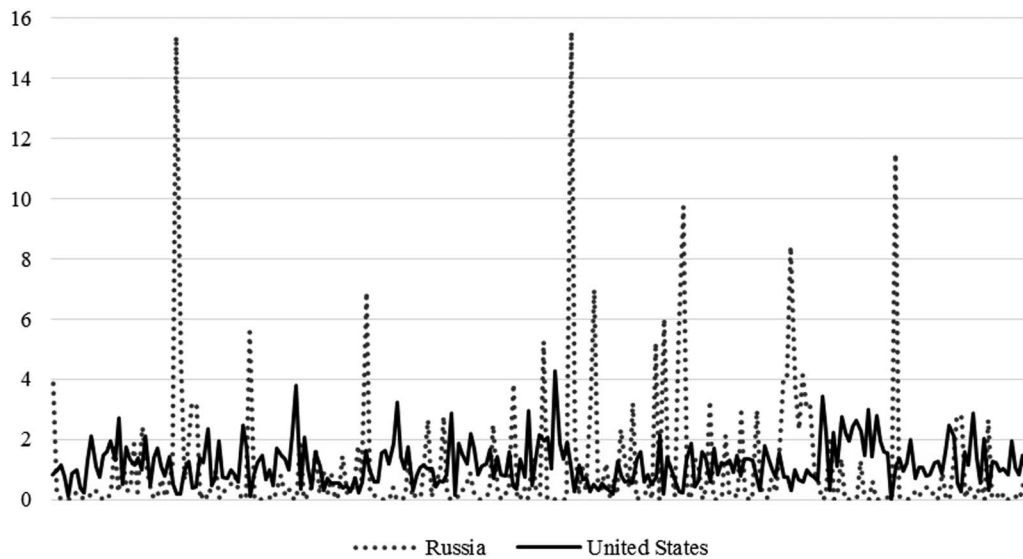
Figure 4. United States and Russia: dispersion of national disciplinary profiles, SI based on input data.

For all 199 countries examined, Figures 6 and 7 show, on input and output sides, the world quantile maps of the GINI coefficient of the SI specialization index. Both maps show the presence of balanced vs. unbalanced research profiles, the former being typical of developed countries, the latter of developing countries. However, not only the "top" countries seen earlier, but almost all (189/199) nations show a higher value of input-based than output-based GINI coefficient (i.e., profiles that are more distributed on the input side). The largest differences are found for Latvia (0.879 vs. 0.650), Luxembourg (0.844 vs. 0.630), and Croatia (0.741 vs. 0.530).
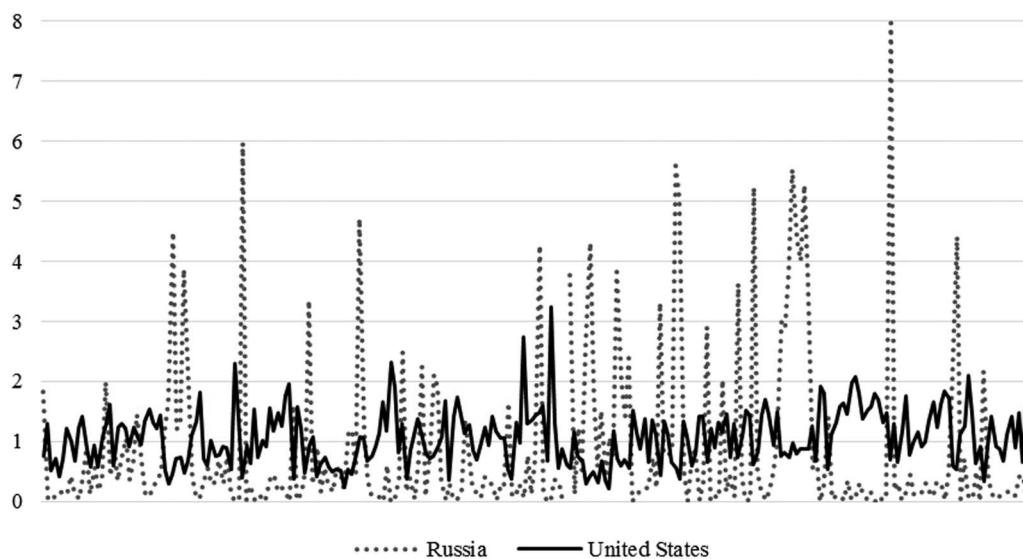
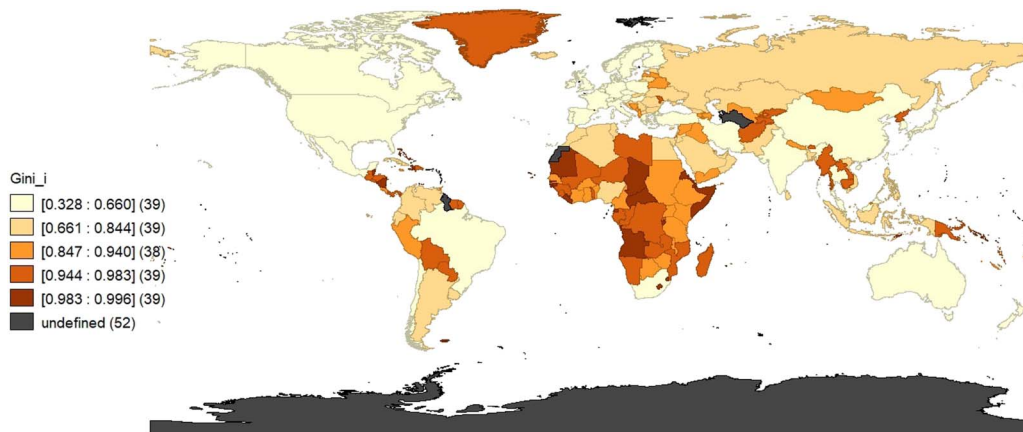**Figure 5.** United States and Russia: dispersion of national disciplinary profiles, SI based on output data.

Figure 6. GINI coefficient of specialization index (SI)—world map based on input.

4.2. Clusters of Countries by Research-System Disciplinary Profile

In the previous sections we used specialization indices based on input and output data to reveal the scientific profile of countries, and especially to compare their disciplinary characterization with respect to all other countries. Such indices can also be used to group countries by similarity of respective profiles. We do this by grouping according to Ward's dissimilarity (Ward, 1963), after principal component analysis (PCA) for reduction of the 254 SC specialization indices to seven principal components[8], beginning from both input and output data. The results are shown Tables 5 and 6, for input and output. There is partial overlapping in the composition of the identified groups but also an evident partial reconfiguration of the clusters when considering one or the other sides of data.

Taking either approach, the first cluster lacks the top countries by share of output seen earlier, including only East African countries, with Ghana also in the output approach.

China, India, and Iran gather in a cluster in both approaches, but the other associated countries change: Taking the input approach, the cluster includes a concentration of Middle Eastern, Asian, and North African countries, united (apart from a few) by linguistic-cultural factors, among which are some "tigers of the East" (Indonesia, Malaysia, Thailand).

Russia occupies a cluster as the sole top country, along with three post-Soviet countries also (Belarus, Kazakhstan, Ukraine). Note that many of the other post-Soviet countries appear in cluster 7 in the input approach, without any top country by share of output; and in cluster 3 in the output approach (along with Poland as a top country).

---

[8] "Principal components" are new variables constructed as linear combinations of initial variables. The initial variables are the SIs on 254 SCs, combined so that the new variables are uncorrelated and most information within the initial variables is stored in the first components. Here, 254-dimensional data yields 254 principal components, but PCA maximizes information in the first ones, achieving a reduced data set focused on the first few components but without important loss of information. Specifically, the first seven components explain about 50% of the variability of the original information, both with input and with output data. Hence, we limit our analysis to these seven components and to as many clusters of countries.
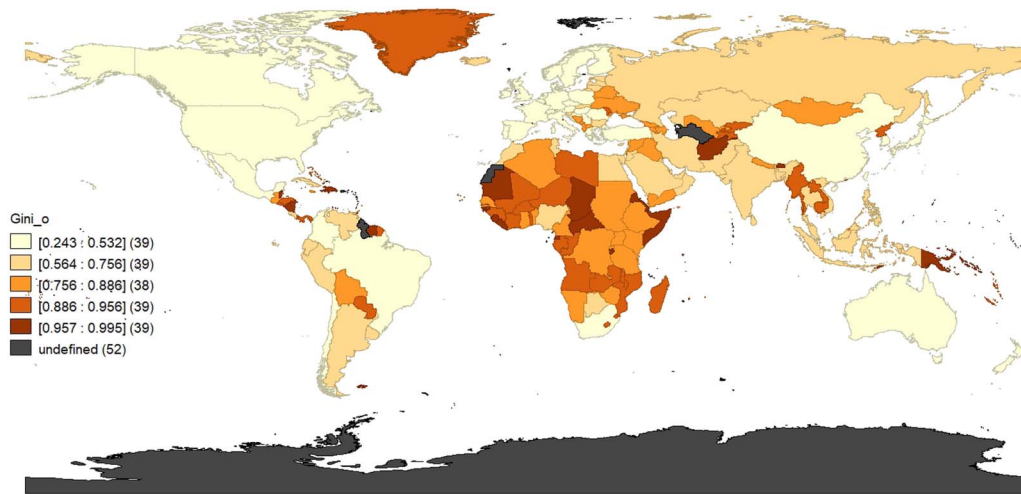
Figure 7. GINI coefficient of specialization index (SI)—world map based on output.

Clusters 5 (input data) and 6 (output data) are quite similar, with the top countries all English-speaking plus the Netherlands in the input approach, and Netherlands plus Sweden in the output approach.

France, Germany, Italy, and Switzerland are all present in clusters 6 (input data) and 7 (output data). Spain groups with these only for the input approach, while considering the output side, it appears as the sole top country of a cluster together with a number of Latin American countries. The situation of Japan is also singular, being associated with Brazil and Poland in the input approach and with France, Germany, Italy, and Switzerland in the output approach.

Table 5. Clustering of countries (based on Ward's dissimilarity), after principal component analysis related to input data, reducing the 254 subject categories specialization indexes to seven principal components

Cluster	Top countries	Other countries
1	–	Ethiopia; Kenya; Tanzania; Uganda
2	Brazil; Japan; Poland	Argentina; Bulgaria; Cameroon; Ecuador; Mexico; Nigeria; Peru; Uruguay; Venezuela
3	China; India; Iran	Algeria; Bangladesh; Colombia; Egypt; Iceland; Indonesia; Iraq; Jordan; Kuwait; Malaysia; Morocco; Oman; Pakistan; Qatar; Romania; Saudi Arabia; Serbia; Sri Lanka; Thailand; Tunisia; United Arab Emirates; Vietnam
4	Russia	Belarus; Kazakhstan; Ukraine
5	Australia; Canada; Netherlands; United Kingdom; United States	Belgium; Ireland; Israel; New Zealand; Norway
6	France; Germany; Italy; South Korea; Spain; Sweden; Switzerland; Turkey	Austria; Chile; Denmark; Finland; Greece; Hungary; Lebanon; Portugal; Singapore; Taiwan
7	–	Croatia; Cyprus; Czech Republic; Estonia; Ghana; Latvia; Lithuania; Luxembourg; Philippines; Slovakia; Slovenia; South Africa

Table 6. Clustering of countries (based on Ward's dissimilarity), after principal component analysis related to output data reducing the 254 subject categories specialization indexes to seven principal components

Cluster	Top countries by share of output	Other countries
1	–	Ethiopia; Ghana; Kenya; Tanzania; Uganda
2	China; India; Iran	Algeria; Egypt; Iraq; Jordan; Luxembourg; Morocco; Pakistan; Qatar; Saudi Arabia; Singapore; Tunisia; United Arab Emirates; Vietnam
3	Brazil; Poland; South Korea; Turkey	Bangladesh; Bulgaria; Cameroon; Croatia; Czech Republic; Greece; Indonesia; Kuwait; Latvia; Lebanon; Lithuania; Malaysia; Nigeria; Oman; Portugal; Romania; Serbia; Slovakia; Slovenia; Sri Lanka; Taiwan; Thailand
4	Russia	Belarus; Kazakhstan; Ukraine
5	Spain	Argentina; Chile; Colombia; Cyprus; Ecuador; Estonia; Iceland; Mexico; Peru; Philippines; South Africa; Uruguay; Venezuela
6	Australia; Canada; Netherlands; Sweden; United Kingdom; United States	Belgium; Denmark; Finland; Ireland; Israel; New Zealand; Norway
7	France; Germany; Italy; Japan; Switzerland	Austria; Hungary

At the same time, with the input data, these four countries correspond to a profile that assimilates that of South Korea and Turkey, countries that instead associate with Brazil and Poland in an output cluster.

Figures 8 and 9 show the ranking of the countries determined by input and output data respectively, but now limiting the analysis solely to principal components 1 and 2: a
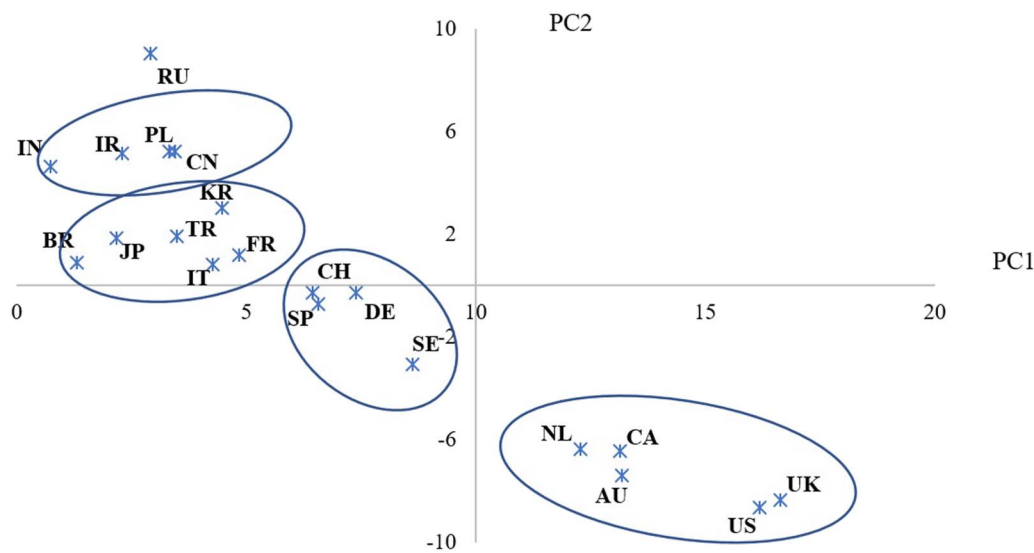

Figure 8. Dispersion of national disciplinary profiles for top 20 countries by share of output, based on the first two principal components related to the input data. AU: Australia; BR: Brazil; CA: Canada; CH: Switzerland; CN: China; DE: Germany; FR: France; IN: India; IR: Iran; IT: Italy; JP: Japan; KR: South Korea; NL: Netherlands; PL: Poland; RU: Russia; SE: Sweden; SP: Spain; TR: Turkey; UK: United Kingdom; US: United States.
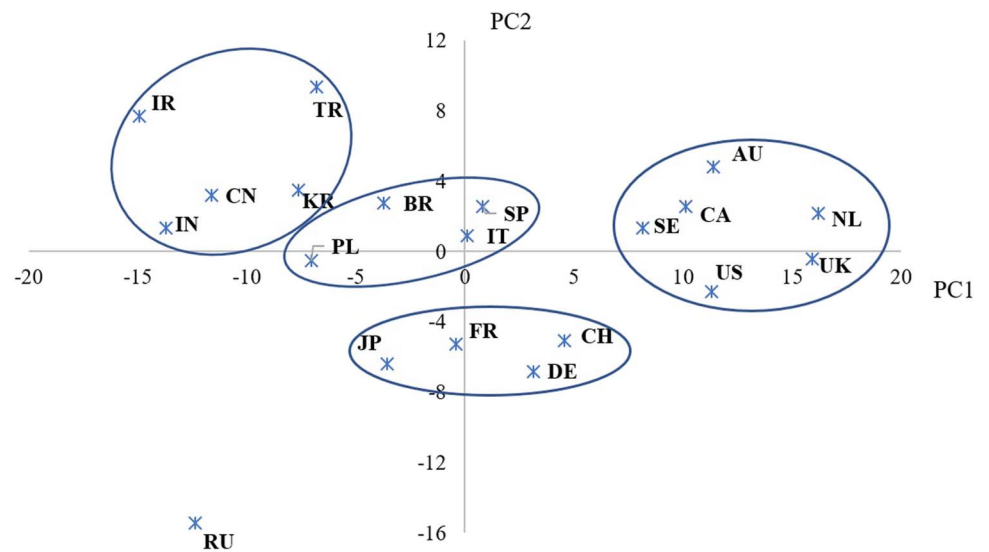
Figure 9. Dispersion of national disciplinary profiles for top 20 countries by share of output, based on the first two principal components related to the output data. AU: Australia; BR: Brazil; CA: Canada; CH: Switzerland; CN: China; DE: Germany; FR: France; IN: India; IR: Iran; IT: Italy; JP: Japan; KR: South Korea; NL: Netherlands; PL: Poland; RU: Russia; SE: Sweden; SP: Spain; TR: Turkey; UK: United Kingdom; US: United States.

representation still more partial on an even greater restriction of the overall information contained in the data[9]. Comparing the two graphs, we see that the rightmost cluster, containing technically and scientifically advanced countries (Australia, Canada, Netherlands, United Kingdom, United States) remains substantially unchanged in composition (with the exception of Sweden, present only for output data), while the other clusters present different recombinations of countries, the only other being the outlier character of Russia, isolated in both graphs.

5. DISCUSSION AND CONCLUSIONS

National research systems can be analyzed in terms of their scientific profiles, and their capital allocation and productive efficiency, through the application of scientific specialization indices (*SI*s), in this way supporting policy-makers as they work to define and pursue the research priorities of their countries. In this paper, we have constructed indices of scientific specialization, calculated from both input and output data, for a set of 199 countries, operating in 254 WoS SCs. One of the aims was to conduct a comparative analysis drawing on the results of the different *SI*s, more specifically: to produce, for each country, a dual specialization profile for each *SC*; for each country and field, to measure the deviations between the values of the two indices; and to observe how distinctive or common features of individual countries or clusters of countries, in terms of their *SI*s for different fields, may vary depending on the point of view of the index used.

For the calculation of the output-based specialization indices, we used the Total Fractional Impact (TFI) (i.e., the sum of the impact of the individual publications produced by the country

---

[9] Note that in Figure 8, PC1 is not centered on zero. The distribution of PC1 is indeed centered on zero for the total 198 countries, but for the 20 largest in our analysis, in the input approach the values are all positive with an average of 6.7.

in each SC). Given that the rate of international collaboration (and therefore coauthorship) in research varies from country to country, we adopted fractional counting to take into account the contribution to each publication by researchers from each country. For calculation of the input-based indices, we used the number of authors from the country in the SC, accepting that, due to lack of information, we could not account for invested capital.

A value above one for *SI* in a given SC indicates a specialization of the country in that SC, evidently because it presents some particular interest. However, based on the construction of the SI as a ratio of ratios, values higher than one are also naturally observed in all those SCs where the share of either TFI or of researchers, although low in value at national level, is nevertheless higher than the corresponding value at world level. This phenomenon is observed for some nationally specific SCs of Art & Humanities, such as "Literature, German, Dutch, Scandinavian" and "Literature, British Isles," for example, where Germany and the United Kingdom are at the top for the relative specialization indices.

Looking at the top 20 countries by share of output, the analysis of their share of SCs presenting differences in indices on output and input sides revealed that most of the G7 countries are characterized by very balanced capital allocation and efficiency of research across fields. Exceptions would be Japan and especially Italy, which falls in a group of opposite character, along with Turkey, Brazil, Poland, and Russia.

On the other hand, the presence of SCs with large shares of the country's total fractional impact or researchers, and with *SI*s much higher than one, is clearly informative of the research system structure, and reflects policy choices that have enhanced the concentration on certain SCs over others.

Depending on the distribution of *SI* values among SCs, a country can therefore have a more or less specialized or diversified disciplinary profile. In this regard, we observed that for all countries but one (Iran), the GINI coefficient for distribution of (*IB*)*SI* is higher than for (*OB*)*SI*. Russia, with the highest values of GINI coefficient on both input and output sides (0.750, 0.706), is the country with the strongest profile of specialization. Russia, along with Iran and India, is also one of the countries with the smallest difference between the two concentration indices: countries that have concentrated most of their resources on only a few sectors, following a historic industrialization model that has accumulated expertise in specific sectors. The contrary profiles of the greatest differences between the (*OB*)*SI* and (*IB*)*SI* are instead seen in Sweden, Switzerland, Canada, and the United States: countries that have diversified their researchers across fields, and which have even more nuanced profiles of specialization when measured through their output.

After PCA, reducing the 254 SC specialization indices to seven principal components, we were able to identify seven clusters of countries by similarities in their profiles. There is partial overlapping in the composition of the identified groups, but also an evident partial reconfiguration of the clusters when considering one or the other sides of data. China, India, and Iran, and four of the English-speaking countries (Australia, Canada, United Kingdom, United States) on the other, compose the nuclei of two groups that maintain similar specialization profiles regardless of the approach.

In concluding, we note that the proposed analysis is not free of the intrinsic limits of the bibliometric approach, inevitably with effects on analytical results. In particular, scientific publications in international scientific journals indexed in WoS represent only part of the total output from research activity. This emerges as a criticality especially where the repertoires provide very low coverage, for example in the fields of Art & Humanities (Aksnes & Sivertsen, 2019), which are fields also suffering from uneven coverage. The choice of field classification

scheme also remains critical. In this work, we implemented the one available in WoS, which covers 254 SCs. The repertoire choice of a high number of fields allows good detail in profiling the specializations of countries, but on the other hand reduces confidence in the analyses, especially for smaller countries.

Other limitations concern citations as a proxy of scholarly impact, as not all citations are positive or indicate real use by citing authors; and citations are not representative of all uses (Abramo, 2018; Bornmann & Daniel, 2008; Tahamtan & Bornmann, 2018; Tahamtan, Safipour Afshar, & Ahamdzadeh, 2016).

Finally, on the input side, the author name disambiguation algorithm is not free of errors, which have an effect also on the accuracy of the output produced by each country. Most importantly, when extracting research staff from publications' metadata, we are not able to account for unproductive researchers or researchers who do not publish in journals indexed in WoS. Furthermore, due to a lack of data on capital investment by country (and even more so by relative fields), the methodological approach to measurement of inputs considers only the numbers of researchers. But research obviously depends on instrumental resources, not only human, and ignoring investment differentials between countries certainly leads to analytical bias. The difference in specialization of a country across fields, from the input and output sides, can in fact have two explanations: higher/lower productivity of the country's researchers but also their higher/lower access to instrumental resources, compared to their colleagues in other countries. For now, the distinction between the two determinants remains difficult to investigate given the lack of data and of a collection framework that is both comprehensive and detailed. On the other hand, however, we are addressing the question of higher/lower differentials in the productivity of researchers by at least examining the feasibility of measurement with respect to an international benchmark, country by country.

DATA AVAILABILITY

Being subject to Clarivate-WoS license restrictions, the raw data cannot be made publicly available. The complete results of our elaborations for all 199 countries in 254 SCs can be found in Abramo et al. (2022b).

REFERENCES

Abramo, G. (2018). Revisiting the scientometric conceptualization of impact and its measurement. *Journal of Informetrics*, *12*(3), 590–597. https://doi.org/10.1016/j.joi.2018.05.001

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012). How important is choice of the scaling factor in standardizing citations? *Journal of Informetrics*, *6*(4), 645–654. https://doi.org/10.1016/j.joi.2012.07.002

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2014). A new bibliometric approach to assess the scientific specialization of regions. *Research Evaluation*, *23*(2), 183–194. https://doi.org/10.1093/reseval/rvu005

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2022a). Revealing the scientific comparative advantage of nations: Common and distinctive features. *Journal of Informetrics*, *16*(1), 101244. https://doi.org/10.1016/j.joi.2021.101244

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2022b). Specialization indexes of countries for 254 subject categories, by input and by output [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.6881520

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, *4*(1), 1–21. https://doi.org/10.2478/jdis-2019-0001

Aksnes, D. W., Sivertsen, G., van Leeuwen, T. N., & Wendt, K. K. (2017). Measuring the productivity of national R&D systems: Challenges in cross-national comparisons of R&D input and publication output indicators. *Science and Public Policy*, *44*(2), 246–258. https://doi.org/10.1093/scipol/scw058

Aksnes, D. W., van Leeuwen, T. N., & Sivertsen, G. (2014). The effect of booming countries on changes in the relative specialization index (RSI) on country level. *Scientometrics*, *101*(2), 1391–1401. https://doi.org/10.1007/s11192-014-1245-3

Allik, J., Realo, A., & Lauk, K. (2020). The scientific impact derived from the disciplinary profiles. *Frontiers in Research Metrics and Analytics*, *5*, 569268. https://doi.org/10.3389/frma.2020.569268, PubMed: 33870047

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. https://doi.org/10.1007/s11192-006-0115-z

Balassa, B. (1965). Trade liberalisation and 'revealed' comparative advantage. *Manchester School of Economic and Social Studies*, *33*(2), 99–123. https://doi.org/10.1111/j.1467-9957.1965.tb00050.x

Bongioanni, I., Daraio, C., Moed, H. F., & Ruocco, G. (2015). Comparing the disciplinary profiles of national and regional research systems by extensive and intensive measures. *Proceedings of ISSI 2015-Istanbul: 15th International Society of Scientometrics and Informetrics Conference* (pp. 684–696).

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Caron, E., & van Eck, N.-J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *Proceedings of the Science and Technology Indicators Conference 2014* (pp. 79–86). Universiteit Leiden.

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*(1), 200–211. https://doi.org/10.1016/j.joi.2016.01.002

Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, *2*(3), 143–148.

Fuchs, J. E., & Heinze, T. (2021). Two-dimensional mapping of university profiles in research. *ISSI2021: 18th International Conference on Scientometrics & Informetrics* (pp. 425–434). KU Leuven, Belgium.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, *31*(121), 124–126. https://doi.org/10.2307/2223319

Glänzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, *48*(2), 121–150. https://doi.org/10.1023/A:1005640604267

Harzing, A.-W., & Giroud, A. (2014). The competitive advantage of nations: An application to academia. *Journal of Informetrics*, *8*(1), 29–42. https://doi.org/10.1016/j.joi.2013.10.007

Heinze, T., Tunger, D., Fuchs, J. E., Jappe, A., & Eberhardt, P. (2019). *Research and teaching profiles of public universities in Germany. A mapping of selected fields*. Wuppertal: BUW.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215. https://doi.org/10.1007/BF02457380

Horta, H., & Veloso, F. M. (2007). Opening the box: Comparing EU and US scientific output by scientific field. *Technological Forecasting and Social Change*, *74*(8), 1334–1356. https://doi.org/10.1016/j.techfore.2007.02.013

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316. https://doi.org/10.1038/430311a, PubMed: 15254529

Leydesdorff, L., & Wagner, C. (2009). Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics*, *3*(4), 353–362. https://doi.org/10.1016/j.joi.2009.05.005

Li, N. (2017). Evolutionary patterns of national disciplinary profiles in research: 1996–2015. *Scientometrics*, *111*(1), 493–520. https://doi.org/10.1007/s11192-017-2259-4

May, R. M. (1997). The scientific wealth of nations. *Science*, *275*, 793–796. https://doi.org/10.1126/science.275.5301.793

Patelli, A., Cimini, G., Pugliese, E., & Gabrielli, A. (2017). The scientific influence of nations on global scientific and technological development. *Journal of Informetrics*, *11*(4), 1229–1237. https://doi.org/10.1016/j.joi.2017.10.005

Rousseau, R. (2018). The *F*-measure for research priority. *Journal of Data and Information Science*, *3*(1), 1–18. https://doi.org/10.2478/jdis-2018-0001

Rousseau, R. (2019). Balassa = revealed competitive advantage = activity. *Scientometrics*, *121*(3), 1835–1836. https://doi.org/10.1007/s11192-019-03273-y

Rousseau, R., & Yang, L. (2012). Reflections on the activity index and related indicators. *Journal of Informetrics*, *6*, 413–421. https://doi.org/10.1016/j.joi.2012.01.004

Sandström, U., & Van den Besselaar, P. (2018). Funding, evaluation, and the performance of national research systems. *Journal of Informetrics*, *12*(1), 365–384. https://doi.org/10.1016/j.joi.2018.01.007

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, *9*(5–6), 281–291. https://doi.org/10.1007/BF02017249

Sugimoto, C. R., & Larivière, V. (2018). *Measuring research: What everyone needs to know.* Oxford: Oxford University Press. https://doi.org/10.1093/wentk/9780190640118.001.0001

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, *12*(1), 203–216. https://doi.org/10.1016/j.joi.2018.01.002

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2

Teixeira, P. N., Rocha, V., Biscaia, R., & Cardoso, M. F. (2012). Competition and diversity in higher education: An empirical approach to specialization patterns of Portuguese institutions. *Higher Education*, *63*(3), 337–352. https://doi.org/10.1007/s10734-011-9444-9

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244. https://doi.org/10.1080/01621459.1963.10500845