Scientometrics
DOI 10.1007/s11192-012-0767-9

A general framework for describing diversity within systems and similarity between systems with applications in informetrics

Qiuju Zhou · Ronald Rousseau · Liying Yang · Ting Yue · Guoliang Yang

Received: 20 October 2011
© Akadémiai Kiadó, Budapest, Hungary 2012

Abstract Building on the ideas of Stirling (J R Soc Interface, 4(15), 707–719, 2007) and Rafols and Meyer (Scientometrics, 82(2), 263–287, 2010), we borrow models of genetic distance based on gene diversity and propose a general conceptual framework to investigate the diversity within and among systems and the similarity between systems. This framework can be used to reveal the relationship of systems weighted by the similarity of the corresponding categories. Application of the framework to scientometrics is explored to evaluate the balance of national disciplinary structures, and the homogeneity of disciplinary structures between countries.

Keywords Diversity · Cosine index · Field-similarity weighted cosine measure · Homogeneity · Interdisciplinarity · Evenness of national disciplinary structure

Introduction

'Diversity' is a concept that features prominently in a variety of disparate disciplines (Stirling 2007). Yet it is almost impossible to find a precise definition. In the nineties,

Q. Zhou (✉) · L. Yang · T. Yue
National Science Library, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
e-mail: zhouqj@mail.las.ac.cn

R. Rousseau
Faculty of Engineering Technology, KHBO (Association KU Leuven), Zeedijk 101, 8400 Oostende, Belgium
e-mail: ronald.rousseau@khbo.be

R. Rousseau
IBW – IOIW, University of Antwerp (UA), Stadscampus, Venusstraat 35, 2000 Antwerpen, Belgium

R. Rousseau
Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3000 Leuven, Belgium

G. Yang
Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

Published online: 26 May 2012

Stirling (1998), following Rao (1982), pointed out that the notion of diversity consists of three basic concepts, namely 'variety', 'evenness' (or balance) and 'disparity' (explained further on). Stirling (2007) presented a general framework for analyzing diversity in science, technology and society.

The field of ecology has a long tradition of studies related to diversity. Consequently, there exists an extensive literature on measures of diversity within populations/communities and dissimilarity or similarity between populations/communities (Gregorius 1987; Lande 1996; Nei 1973; Nei 1978; Rao 1982; Whittaker 1960; Whittaker 1972). Based on the Simpson measure of genetic diversity Nei (1973) developed an analogous measure of genetic similarity [Nei's minimum genetic distance and Nei's standard genetic distance (Nei 1972)] among populations, which has been widely used to describe the genetic structure of populations. Simpson's diversity measure, equal to 1—Simpson's concentration measure (Simpson 1949), is part of Nei's approach and is meant to take two properties into account, namely 'variety' and 'evenness'. Note though that we have shown that this is not quite true, see (Nijssen et al. 1998; Rousseau et al. 1999). Further, Shriver et al. (1995) proposed a measure of genetic distance, based on diversity and generalizing Rao's (1982) approach. Yet there is scant literature, especially in scientometrics and informetrics related to a general framework for analyzing the similarity among systems based on a diversity measure that takes all three properties into account.

We observe that over the years many scientometrics studies have been published about interdisciplinarity, see e.g. (Ackoff 1960; Qin et al. 1997; Porter and Chubin 1985). Moreover, attention has been paid to the issue of the homogeneity of disciplinary structures between countries, see e.g. (Kozlowski et al. 1999; Segev et al. 2007; Zitt et al. 2007). Yet, few studies combine the two topics.

The objectives of this paper are threefold.

(1) Inspired by Stirling (2007) and related ecological research (Nei and Li 1979; Nei and Tajima 1981; Shriver et al. 1995) on the relationship of diversity within and among populations and the similarity between populations, we propose a framework to explore the relationship of diversity within systems and the similarity among systems. This framework can be used to reveal the relationship of systems based on the weighted similarity of categories.
(2) We apply this approach to the field of scientometrics to analyze the balance of national disciplinary structures and the heterogeneity or homogeneity of disciplinary structures between countries. When measuring the diversity and the similarity of national disciplines, we incorporate difference (or similarity) between disciplines in the spirit of Rafols and Meyer (2010).
(3) Finally, we clarify the meaning of the resulting multidimensional scaling (MDS) map using the relative contribution of 32 countries/regions in 14 factors (the result of a factor analysis for 172 ISI subject categories) and 3 major fields.

The paper is further organized as follows. Section 2 presents the concepts of diversity and similarity, and relates them to the literature on interdisciplinarity and national disciplinary structures. Section 3 introduces a general conceptual framework to investigate the diversity within and the similarity among systems. It is indicated how this framework can be applied in informetrics. Section 4 provides a case study related to national disciplinary structures and discusses the results. Section 5 focuses on the difference between the Simpson diversity measure and the Stirling–Rao diversity measure. Section 6 concludes by summarizing the results and discussing their implications. Finally the appendices contain some additional information.

Review of related studies

The concept of diversity

Diversity is an attribute of any system whose elements can be divided into categories (Leonard and Jones 1989). It can be described as the property of how the elements of a system are apportioned into categories (Stirling 2007). The discipline of *mathematical ecology* is an area in which diversity concepts have been thoroughly investigated (Magurran 1988). Stirling (1998) pointed out that diversity concepts have three basic properties, namely 'variety', 'evenness' (balance), and 'disparity'. We recall from Stirling (2007) the main aspects of these three basic notions:

- Variety is the number of categories into which a system is divided. This aspect of diversity is highlighted, for instance, in the use of species-number indices in ecology (McIntosh 1967). All else being equal, the greater the variety, the greater the diversity.
- Evenness (balance) is a function of the pattern of apportionment of elements across categories. Evenness or balance in the distribution of categories is referred to as 'evenness' in ecology (Pielou 1969; Nijssen et al. 1998) and 'concentration' in economics (Finkelstein and Friedman 1967). Also in the information sciences the terms 'concentration' and 'inequality' are more popular than 'evenness' (Egghe and Rousseau 1991; Rousseau 1992; Burrell 2005). For fixed variety this notion is captured by the (diversity version) of the Shannon–Wiener (also called Theil), Gini and Simpson indices (Magurran 1988). All else being equal, the larger the evenness, the greater the diversity.
- Disparity (similarity) is the degree to which categories are different (similar). Disparity refers to the manner and degree in which categories may be distinguished (Runnegar 1987). This is addressed by an array of taxonomic indices in paleontology (Williams and Humphries 1994), conservation biology (Solow and Polasky 1994) and economics (Nguyen et al. 2005). This notion is usually based on some 'distance' measure. All else being equal, the more disparate the represented elements, the greater the diversity.

Each of these three aspects separately is a necessary but insufficient property of diversity (Clarke 1978; Sokal and Sneath 1970; Rousseau and Van Hecke 1999; Stirling 2007). Although addressed by different names, each is applied across a range of disciplines and aggregated in various ways to set up quantitative indices (Hill 1973; Patil and Taillie 1982). Stirling (1998, 2007) provided a framework and gave general formulae to measure the diversity within systems, based on the three basic properties of diversity. We will show that his approach is applicable to the study of multidisciplinary research.

Study of disciplinary diversity in informetrics

Morillo et al. (2003) investigated interdisciplinarity and addressed properties of diversity, namely variety, evenness, and similarity which they based on the Thomson Reuters—ISI multi-assignation of journals to subject categories (SCs).

Some studies combined the properties of variety and evenness into one indicator, such as the Pratt index (Morillo et al. 2001) or the Shannon entropy (Barjak and Robinson 2008). Porter et al. (2007), Rafols and Meyer (2007, 2010) actually integrated the three attributes of variety, evenness and similarity into one indicator. These authors introduced Stirling's diversity into scientometrics as a disciplinary diversity indicator of interdisciplinarity to describe the heterogeneity of a bibliometric set viewed from predefined

categories, i.e., using a top-down approach that locates the set on the global map of science. Interestingly, Porter's indicator of integration was a particular parameterization of Stirling's index, where the similarities S_{ij} are the cosine measure for co-citation patterns among ISI SCs. Barjak and Robinson (2009) also applied an integration of the three aspects of diversity.

The concept of similarity

In ecology, there exists an extensive literature on measures of diversity within populations/communities and dissimilarity or similarity between populations/communities (Nei 1978; Rao 1982). Gregorius (1987) discussed the relationship between the concepts of genetic diversity and differentiation. Rao (1982) proposed a unified approach for measuring diversity and dissimilarity coefficients. Two commonly used genetic distance measures, Nei's minimum genetic distance and Nei's standard genetic distance (Nei 1972), are based on genetic diversity within and between populations.

Similarity of disciplines

Most informetric studies related to the similarity between disciplines or SCs can be divided into two groups: those for which the focal point is mapping and those focusing on interdisciplinarity. Various studies of interdisciplinarity have been based on the assumption that journals can be grouped using the ISI SCs (e.g., van Leeuwen and Tijssen 2000; Morillo et al. 2001, 2003).

"Interdisciplinarity" is often a policy objective because it is believed that new developments frequently take place at the borders of disciplines (Zitt 2005). One of the potential uses of a map of science is to help us understand the cognitive base and the relative positions of emerging fields (Bordons et al. 2004; Porter et al. 2007; van Raan and van Leeuwen 2002). Leydesdorff (2007) explored network centralities as indicators of interdisciplinarity. Liu et al. (2012) saw interdisciplinarity as one side of the knowledge integration (via interdisciplinary work) and knowledge diffusion duality and proposed a general framework for their study.

In an informetric context linkages between categories, especially those shown through citation relations, can be regarded as tokens of cognitive similarity. The matrix of citation relations among categories is often used to map the backbone of science (Leydesdorff and Rafols 2009; Zitt et al. 2005). In their studies on interdisciplinarity Rafols and Meyer (2007, 2010) used bibliographic coupling strength as the degree of similarity between ISI SCs. Morillo et al. (2003) provided another way to establish the relationships between two given categories based on multi-assignation of journals by ISI. Yet, there are few studies exploring the relationship between the concepts of diversity and dissimilarity or similarity of systems.

National disciplinary structures

In scientometric studies national disciplinary structures can be investigated via publication profiles (Glänzel 2001; Glänzel et al. 2008; Hu and Rousseau 2009). These profiles reveal interesting aspects of a country's research landscape and its strategies and preferences in science policy and organization (Zhang et al. 2011). The disciplinary structure of science

for individual countries determines how resources and funds are distributed over disciplines and are influenced by culture, political history and geography (Almeida et al. 2009).

Radar plots derived from the Activity Index (Frame 1977) and the relative specialization index (Glänzel 2001) are often used to represent the balance in national disciplinary structures (King 2004; Yue 2008). By comparing the structure of these radar plots, the homogeneity and heterogeneity of national disciplinary structure can be distinguished. However, the structure of radar plots does not directly provide an index to quantify disciplinary balance. Homogeneity and heterogeneity of national disciplinary structures are often described visually. However, some scientists did use quantitative measures of homogeneity for national disciplinary structures. Notably, Liang et al. (2006) discussed structural similarities between the Chinese science system and that of Western systems using the approach of Stanley et al. (1996), by considering the growth dynamics of Chinese universities and journals. Kozlowski et al. (1999) explored the homogeneity of disciplinary structure between countries based on 102 ISI disciplines using the cosine index, while Yue et al. (2011) compared the similarity and balance of the disciplinary structure in science between the G7 and the BRIC countries. However, in all these studies related to the homogeneity or the heterogeneity of disciplinary structures, disciplines are assumed to be independent. In other words, similarity or dissimilarity of disciplines is not taken into account. This is the main aspect we want to introduce in this paper.

Conceptual framework

General framework and formulae

Consider a two-way N × M cross-classification table. The M columns will be referred to as systems, while the N rows are referred to as categories. In our application(s) systems will be countries, categories will be SCI SCs (in short SCs), used to operationalize the notion of scientific fields. Data in the table are the numbers of publications per country per subject category.

Our aim is to determine the similarity (according to a given similarity measure) between systems (countries) based on the diversity over categories per system (diversity over fields per country).

As one of the requirements for acceptable diversity measures is scale invariance (Dalton 1920), we do not need the actual number of publications per SC, per country, but only the relative number of publications per SC. The term 'relative' means: with respect to the country under study. Table 1 provides an example of such a two-way cross-classification table. The proportion of items (publications) in system X and category j is denoted as p_{Xj}.

We note that in our investigation the number of categories (number of species) is fixed and set equal to N. Hence basic diversity can be measured through evenness, while the full

Table 1 An illustrative table of systems, categories and proportions

	System 1	System 2	System 3
Category 1	0.8	0.65	0.2
Category 2	0.05	0	0.35
Category 3	0.15	0.05	0.25
Category 4	0	0.3	0.2
Total	1	1	1

(Stirling–Rao type) diversity is measured through evenness and disparity. Disparity between categories i and j is denoted as D_{ij} (with $0 \leq D_{ij} = D_{ji} \leq 1$), leading to a square N × N symmetric disparity matrix D. Its diagonal values, D_{ii}, are equal to 0. Its dual notion, similarity between categories, is denoted as S_{ij} and defined as $S_{ij} = 1 - D_{ij}$. All S_{ij}-values lead to an N × N similarity matrix S. Table 2 provides an example of such a square matrix. Note that there are no other requirements for its elements besides $0 \leq S_{ij} = S_{\mathrm{ji}} \leq 1$ and $S_{ii} = 1$ ($D_{ii} = 0$).

Inspired by Nei and Li (1979), Nei and Tajima (1981), Rao (1982), Shriver et al. (1995) and Stirling (2007) we use Simpson's diversity measure for basic diversity and its similarity-weighted form (see further for its exact definition) for the full diversity. We recall that one minus the Simpson diversity index is known as the Simpson concentration index. In the economic literature this concentration index is usually called the Herfindahl or Herfindahl–Hirschman index.

For a given system X with N categories the Simpson diversity, denoted as L_{X}, is calculated as:

$$L_{\mathrm{X}} = 1 - \sum_{i=1}^{N} p_{\mathrm{X}_i}^2 \tag{1}$$

While Simpson's concentration measure (Simpson, 1949) is:

$$\lambda_{\mathrm{X}} = \sum_{i=1}^{N} p_{\mathrm{X}i}^2 \tag{2}$$

The general Stirling–Rao diversity, denoted as SR_{X}, for this same system is calculated as:

$$SR_{\mathrm{X}} = \sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} D_{ij} = \sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} (1 - S_{ij}) \tag{3}$$

$$= 2 \sum_{i=1}^{N} \left(\sum_{j=i+1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} D_{ij} \right) = 2 \sum_{i=1}^{N} \left(\sum_{j=i+1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} (1 - S_{ij}) \right) \tag{4}$$

It can be shown, see Appendix A, that Eq. 3 is also equal to

$$SR_{\mathrm{X}} = \sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} D_{ij} = 1 - \sum_{i,j=1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} S_{ij} \tag{5}$$

By analogy we define the Stirling–Rao concentration measure of system X as:

$$\varphi_{\mathrm{X}} = \sum_{i=1}^{N} \left(\sum_{j=i+1}^{N} p_{\mathrm{X}_i} p_{\mathrm{X}_j} S_{ij} \right) \tag{6}$$

Table 2 An example of a similarity matrix S for the categories shown in Table 1

	Category 1	Category 2	Category 3	Category 4
Category 1	1	0.95	0.07	0.08
Category 2	0.95	1	0.09	0.02
Category 3	0.07	0.09	1	0.96
Category 4	0.08	0.02	0.96	1

Table 3 Simpson and Stirling–Rao diversity values based on the systems and categories values shown in Tables 1 and 2

	System 1	System 2	System 3
Simpson diversity	0.335	0.485	0.735
Stirling–Rao diversity	0.241	0.420	0.474

Clearly L_X can be derived from SR_X by taking the matrix S equal to the N × N unit matrix I_N (see Appendix A). The corresponding D matrix is the N × N matrix in which all entries are 1 except for the diagonal ones which are zero. Note that the larger the similarity values S_{ij} (or the smaller the disparity values D_{ij}), the smaller the diversity, which is a logical requirement.

Based on Tables 1 and 2, Table 3 shows the corresponding Simpson and Stirling–Rao diversity values.

We see from the examples in Table 3 that the Stirling–Rao diversity is always smaller than the Simpson diversity. This can easily be shown in general as follows:

$$SR_X = \sum_{i,j=1}^{N} p_{X_i} p_{X_j} D_{ij} = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} p_{X_i} p_{X_j} D_{ij} \leq \sum_{\substack{i,j=1 \\ i \neq j}}^{N} p_{X_i} p_{X_j} = 1 - \sum_{i=1}^{N} p_{X_i}^2 = L_X \tag{7}$$

Where we have used the fact that for each i $D_{ii} = 0$. The inequality sign comes from the fact that $D_{ij} \leq 1$, and the validity of the penultimate equality is shown in Appendix A.

In the same vein similarity between systems X and Y (with the same number of categories N), denoted as $\varphi_{X,Y}$, is defined as:

$$\varphi_{X,Y} = \sum_{i,j=1}^{N} p_{X_i} p_{Y_j} S_{ij} \tag{8}$$

Its elementary form with $S = I_N$ leads to the expression:

$$\lambda_{X,Y} = \sum_{i=1}^{N} p_{X_i} p_{Y_i} \tag{9}$$

In order to understand the term 'similarity' in this context we consider the normalized form of Eq. 9 first. This normalized form is defined as:

$$\lambda(X,Y) = \frac{\lambda_{X,Y}}{\sqrt{\lambda_X . \lambda_Y}} = \frac{\sum_{i=1}^{N} p_{X_i} p_{Y_i}}{\sqrt{\left(\sum_{i=1}^{N} p_{X_i}^2\right)\left(\sum_{i=1}^{N} p_{Y_i}^2\right)}} \tag{10}$$

Clearly, λ is nothing but the cosine or Salton similarity measure of the arrays $(p_X)_{i=1,\ldots,N}$ and $(p_Y)_{i=1,\ldots N}$. Similarly we consider the normalized form:

$$\varphi(X,Y) = \frac{\varphi_{X,Y}}{\sqrt{\varphi_X \varphi_Y}} = \frac{\sum_{i,j=1}^{N} p_{X_i} p_{Y_j} S_{ij}}{\sqrt{\left(\sum_{i,j=1}^{N} p_{X_i} p_{X_j} S_{ij}\right) . \left(\sum_{i,j=1}^{N} p_{Y_i} p_{Y_j} S_{ij}\right)}} \tag{11}$$

This normalized form can be considered a weighted cosine measure based on similarities between categories. We will refer to it as the similarity-weighted cosine measure. It leads to a similarity value between systems. Tables 4 and 5 show cosine similarity values and similarity-weighted cosine values for the examples introduced in Tables 1 and 2.

Table 4 Cosine similarity values

	System 1	System 2	System 3
System 1	1	0.901	0.512
System 2		1	0.548
System 3			1

Table 5 Similarity-weighted cosine values

	System 1	System 2	System 3
System 1	1	0.951	0.877
System 2		1	0.962
System 3			1

Ranking systems according to similarity we find for the cosine similarity the ranking (from most similar to least):

(system 1 · system 2) > (system 2 · system 3) > (system 1 · system 3)

While according to the similarity-weighted cosine measure we find:

(system 2 · system 3) > (system 1 · system 2) > (system 1 · system 3)

We see in this example that taking similarity between categories into account not only changes the ranking, but also the relative values of system similarities.

We note that some important mathematical considerations must be mentioned in relation to the use of Eq. 11. These are explained in Appendix A.

An overview of the conceptual framework

We pause for a moment and explain what we have introduced so far. Diversity within systems can be described using the number of categories (number of species in ecology; scientific fields in scientometric studies) and evenness; or one can go one step further and include a measure for the disparity between categories. In the first case one may use a classical measure such as the Gini evenness index or the Simpson diversity measure, in the more general case the measure proposed in the literature is the Stirling–Rao diversity measure.

Similarity between systems can be studied by using a classical similarity measure such as Salton's cosine measure or a weighted form taking category similarity into account. For this purpose we propose the similarity-weighted cosine measure as given by Eq. 11. All this is recapitulated in Table 6.

This framework's applicability in informetrics

Rafols and Meyer (2007, 2010) introduced the Stirling diversity in scientometrics as an indicator of interdisciplinarity. They described the heterogeneity of a bibliometric set viewed from predefined categories, namely the ISI SCs, where the similarities S_{ij} are Salton's cosine values for co-citation patterns among ISI SCs.

Within the Essential Science Indicators (ESI) the science system is divided into 22 broad fields in such a way that every journal belongs to exactly one field. It is implied that

Table 6 Overview of the proposed framework for the study of diversity and similarity

Diversity within systems		Diversity between systems	
With a given number of independent categories in the system		With a given number of independent categories in the system	
We use:	Simpson diversity measure, Eq. 1	We use:	Salton's cosine measure, Eq. 10
Taking disparity (or similarity) between categories into account		Taking disparity (or similarity) between categories into account	
We use:	Stirling–Rao measure, Eqs. 3–5	We use:	Similarity-weighted cosine measure, Eq. 11

the 22 broad fields are independent. Unlike the 22 broad fields, the ISI SCs are not disjointedly or hierarchically organized, but interconnected, because more than one category is often attributed to a journal (Leydesdorff and Rafols 2009). Furthermore, these are more specific and therefore contain more information. As we use a method that is designed especially to take similarity between categories into account, we prefer the more detailed approach with more than 170 subdivisions.

In our study of national disciplinary structure we can consider a country to be a system and the ISI articles to be elements. Then, the articles can be divided into different ISI SCs. The framework of diversity and similarity can be used to investigate the balance of the national disciplinary structure of a certain country, and the homogeneity of a group of countries, and to extract a pattern of publication profiles.

Kozlowski et al. (1999) explored the homogeneity of disciplinary structure between countries based on 102 ISI disciplines by using the Cosine Index. Yue et al. (2011) compared the disciplinary structure in science between G7 and the BRIC countries based on 170 + ISI SCs, using the cosine index as similarity measure. Because of the methods used these studies assume independence between disciplines. In terms of the notation introduced above this means that they took, for $i \neq j$: $D_{ij} = 1$ or $S_{ij} = 0$. Yet, in reality ISI categories are not independent. Hence, as argued further on, we think that Kozlowski et al. (1999) and Yue et al. (2011) underestimated the similarity between countries (but see Appendix A).

We should note that the idea of measuring similarity between systems has been proposed in a slightly different setting by Burrell (2005, 2006). His systems are general informetric distributions for which the numbers of categories are not necessarily the same. He not only performed a theoretical study based on continuous distribution functions but also considered the discrete case. This led to a relative concentration coefficient and a coefficient of co-concentration, based on the Gini index and derived from Dagum's notion of relative economic affluence (Dagum 1987). He, however, did not take disparity into account.

In the following sections we show the results of an empirical study. To refine similarity results based on the cosine index we will use the similarity-weighted cosine index. Diversity within countries will be measured using the Stirling–Rao index. An investigation of the similarities between countries leads to a general view on the homogeneity (or heterogeneity) of the group of countries under study.

A case study: purpose, data and methods

Research purpose of the case study

In this case study we have the following four goals

- Illustrate the use of the framework introduced in the first part of this article;
- Investigate if the cosine similarity and the similarity-weighted cosine similarity lead to different similarity results between systems;
- Try to explain observed differences;
- Investigate if the Simpson and the Stirling–Rao diversity measures lead to different diversity values and/or rankings within systems.

Data

Data are extracted from the National Science Indicators (Thomson Reuters) database. We chose 32 countries including the G8 group and the 15 countries of the European Union before the 2004 accession (EU15). Concretely, the following countries are included in our case study: Australia, Austria, Belgium, Brazil, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, India, Iran, Israel, Italy, Japan, Mexico, the Netherlands, Poland, Portugal, Russia, Singapore, Spain, South Africa, South Korea, Sweden, Switzerland, Taiwan, Turkey, the United Kingdom and the United States. The publications in 2009 of each country were counted and grouped into 172 SCI subject categories. As categories are scientific fields we use a field-similarity weighted cosine measure. In order to obtain the Stirling–Rao diversity index, a field-similarity matrix S_{ij} in the cited dimension provided by Leydesdorff and Rafols (2009) was chosen (see Table 7). The S_{ij} describes the similarity in the citation patterns for each pair of SCs in 2006. We note that this S-matrix takes two related but different phenomena into account: one is the overlapping between SCs, and the other is the content similarity between SCs.

Calculations

The proportions (p_{xi}) of 172 SCI subject categories in each country were computed (see Table 8). By combining the SC proportions (p_{xi}) and the similarities S_{ij}, we computed the

Table 7 Similarity matrix S_{ij} of 172 SCI subject categories—partim

Number	172 SCs	1	2	3	4	5	6	7	.....
1	Biochemistry & Molecular Biology	1.0000	0.9760	0.9489	0.0226	0.3372	0.1406	0.0157	.....
2	Biophysics	0.9760	1.0000	0.9041	0.0375	0.3444	0.1947	0.0259	.....
3	Cell Biology	0.9489	0.9041	1.0000	0.0126	0.2492	0.0928	0.0130	.....
4	Thermodynamics	0.0226	0.0375	0.0126	1.0000	0.1738	0.2883	0.7688	.....
5	Chemistry, Applied	0.3372	0.3444	0.2492	0.1738	1.0000	0.5053	0.0865	.....
6	Chemistry, Physical	0.1406	0.1947	0.0928	0.2883	0.5053	1.0000	0.1459	.....
7	Engineering, Mechanical	0.0157	0.0259	0.0130	0.7688	0.0865	0.1459	1.0000	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Table 8 SC Proportions (p_{xi}) of 32 countries/regions (expressed in %)—partim

172 SCs	USA	China	Japan	Germany	UK	France	Italy	.....
Ornithology	0.07	0.01	0.00	0.02	0.08	0.03	0.01	.....
Computer Science, Software Engineering	0.35	0.37	0.22	0.25	0.33	0.32	0.31	.....
Automation and Control Systems	0.17	0.62	0.26	0.22	0.23	0.35	0.32	.....
Entomology	0.32	0.18	0.27	0.13	0.16	0.14	0.17	.....
Infectious Diseases	0.70	0.20	0.36	0.48	1.08	0.92	0.84	.....
Public, Environmental and Occupational Health	1.47	0.23	0.27	0.49	1.25	0.78	0.63	.....
Nanoscience and Nanotechnology	0.79	1.66	0.96	0.80	0.54	0.79	0.58	.....
Nursing	0.62	0.04	0.04	0.05	0.43	0.01	0.05	.....
Engineering, Biomedical	0.48	0.34	0.41	0.46	0.44	0.49	0.60	.....
Electrochemistry	0.20	0.91	0.58	0.27	0.19	0.35	0.36	.....
Materials Science, Composites	0.07	0.18	0.08	0.11	0.11	0.15	0.09	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....

values of the indicators given by Eqs. 1–11 and made comparisons across the 32 countries/regions mentioned above.

Methods

Indicator values were calculated in Matlab. Stirling–Rao diversity of each country was calculated to obtain the evenness of a country's disciplinary structure. To study the homogeneity of the disciplinary structure among different countries/regions, a similarity matrix based on the normalized *φ-Index* (Eq. 11) was used to perform cluster and MDS analyses. Cluster analysis, using Ward's Method, was performed with SPSS, while MDS analysis was performed with Ucinet 6.0, choosing a similarity matrix as input for the two procedures. We note that Ward's minimum variance method for hierarchical clustering is generally considered to give good results (and certainly better than the single link method), see e.g. (Griffiths et al. 1984).

Leydesdorff and Rafols (2009) applied factor analysis based on the citation matrix of 172 SCs, suggesting a 14-factor solution with a minimum loss of information. Factor analysis is a family of multivariate statistical techniques for examining correlations amongst variables. It is used to identify clusters of inter-correlated variables (here meaning SCs). Factors consist of relatively homogeneous variables. That is to say factor analysis taking category similarity into account. Considering this we confirm that Leydesdorff's 14 disciplines can interpret the results of the field similarity-weighted cosine similarity.

We use Leydesdorff's 14 disciplines clustered further into 3 major fields: Life Sciences, Physical Sciences and Environmental Sciences. The proportion of the 32 countries' articles in these 14 disciplines and 3 major fields are used to give an interpretation of Figs. 1 and 2, illustrating the homogeneity (or heterogeneity) of the disciplinary structure of science.

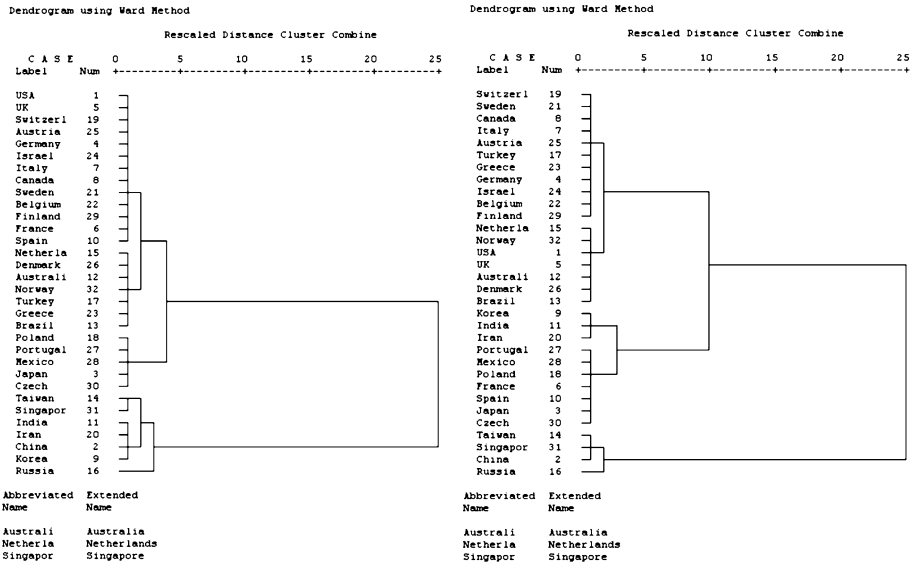


Fig. 1 Dendrograms using Ward's clustering analysis of national disciplinary structure. *Left* based on the cosine index, *Right* based on the similarity-weighted cosine index

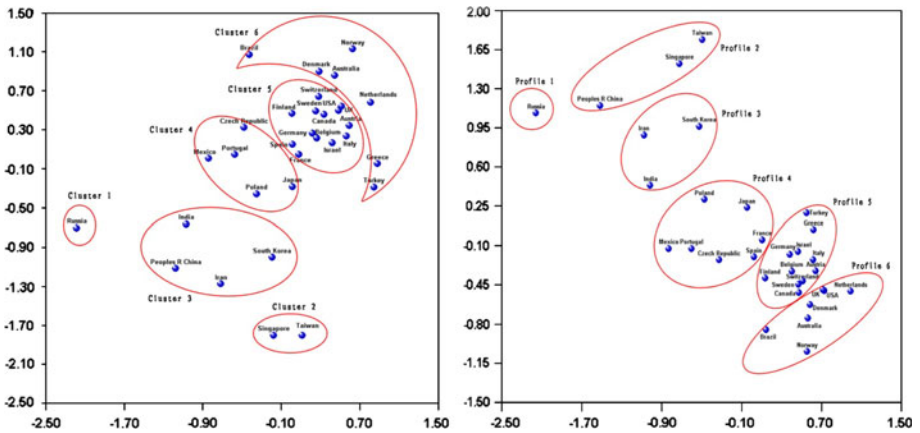


Fig. 2 Maps resulting from MDS analysis of national disciplinary structure. *Left* based on the cosine index, *Right* based on the similarity-weighted cosine index

Results and discussion

A comparison between the results obtained with the cosine and with the field-similarity weighted cosine index

Figure 1 shows the final dendrograms using Ward's Method based on the cosine index and field-similarity weighted cosine index. Figure 2 shows the results of the MDS analysis. The proportion of the countries/regions' articles in the 14 disciplines and 3 major fields are used to explain Figs. 1 and 2.

Since Ward's method is a bottom-up agglomerative clustering method (each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy) we compare the results of the two methods in a "bottom-up" fashion.

The first levels of the two dendrograms show that the 32 countries/regions were grouped into six different homogeneous clusters. Countries belonging to the same cluster are considered to have a rather homogeneous disciplinary structure, while countries in different clusters are more heterogeneous. Yet, Figs. 1 and 2 illustrate the fact that the two measures lead to different results. Indeed, although we have six clusters in each case, their members are not the same. Based on the cosine index we have the following six clusters:

Cluster 1: Russia
Cluster 2: Singapore, Taiwan
Cluster 3: China, India, Iran, South Korea
Cluster 4: Czech Republic, Japan, Mexico, Poland, Portugal
Cluster 5: Austria, Belgium, Canada, Finland, France, Germany, Israel, Italy, Spain, Sweden, Switzerland, UK, USA
Cluster 6: Australia, Brazil, Denmark, Greece, the Netherlands, Norway, Turkey

The field-similarity weighted cosine index leads to the following six clusters, to which we refer as 'profiles':

Profile 1: Russia
Profile 2: China, Singapore, Taiwan
Profile 3: India, Iran, South Korea
Profile 4: Czech Republic, France, Japan, Mexico, Poland, Portugal
Profile 5: Austria, Belgium, Canada, Germany, Greece, Israel, Italy, Sweden, Switzerland, Turkey
Profile 6: Australia, Brazil, Denmark, the Netherlands, Norway, UK, USA

We see that, for instance, China moves away from Iran, India and South Korea (Cluster 3) and joins Singapore and Taiwan in Profile 2. Next, as an example, we consider the case of Turkey and Greece in more detail. We can see that based on the two measures Turkey and Greece always belong to the same cluster. Yet, in the MDS analysis based on the cosine index (Fig. 3), they belong to the same cluster as Australia, Brazil, Denmark, the Netherlands and Norway. Taking field-similarity into account, places Turkey and Greece in Profile 5 together with many West European countries, while Australia, Denmark, Norway, Brazil and the Netherlands are now in Profile 6 with the UK and the USA.

Next we discuss these findings from another perspective. We consider again Greece and Turkey, and compare with the Netherlands. The columns of Table 9 show for these three countries the percentages of articles in each of the 14 disciplines and each of the major fields. Percentages should sum to 100 % but because of rounding errors this is not always the case. We define *Top disciplines* for a country those disciplines in which a country publishes 10 % or more of its total scientific output. These are indicated in bold italics. These three countries have the same number of *Top disciplines*. Moreover all three have Biomedical Sciences and Clinical Sciences among their *Top disciplines*. For Greece and Turkey the third *Top discipline* is in the Physical Sciences, while for the Netherlands the third *Top discipline* is in the Life Sciences. MDS based on the similarity-weighted cosine index captures this difference. In view of this we use Stirling–Rao diversity to determine the balance of the disciplinary structure of the 32 countries/regions and the field-similarity weighted cosine measure to describe the homogeneity of the disciplinary structure of countries.

Table 9 Details for three countries according to 14 disciplines and 3 major fields (percentages of articles in each discipline or major field)

	172 SCs	Turkey	Greece	The Netherlands
14 Disciplines	Biomedical Sciences	***17***	***23***	***23***
	Clinical Sciences	***19***	***18***	***23***
	Neuro-Sciences	6	5	***11***
	Infectious Diseases	6	5	6
	Gen. Medicine; Health	4	3	5
	Materials Sciences	7	8	5
	Chemistry	***10***	8	6
	Computer Sciences	8	***10***	5
	Engineering	6	4	3
	Physics	4	4	3
	Ecology	2	2	3
	Environmental Sciences	6	5	3
	Geo-sciences	2	3	2
	Agriculture	4	3	2
3 Major fields	Life Sciences	52	53	68
	Physical Sciences	35	34	22
	Environmental Sciences	14	13	10

An interpretation of the six profiles from the perspective of average proportions of articles in the 14 disciplines and 3 major fields

We note that the six profiles, see Table 10, are ranked from left to right in ascending order of contributions to the Life Sciences and in descending order of contributions to the Physical Sciences. When considering three major fields we see that Profile 4 has the smallest variability, while Profile 1 has by far the largest. Hence research is most polarized in Profile 1 (Russia) and least in Profile 4 (Czech Republic, France, Japan, Mexico, Poland, Portugal). For the other four profiles we may say that in Profiles 5 and 6 the Life Sciences dominate, while for the other two (Profiles 2 and 3) the Physical Sciences dominate. These conclusions are confirmed when taking so-called *Top disciplines* into account. Recall that we have defined *Top disciplines* as those disciplines in which a country published at least 10 % of its articles.

In Profile 1, Russia, the four *Top disciplines* are Materials Sciences, Chemistry, Physics and Biomedical Sciences. The former three disciplines belong to the Physical Sciences, and the last one in Life Sciences. For the Profile 2, including China, one may say that Materials Sciences, Computer Sciences, Chemistry, and Biomedical Sciences dominate, i.e. 3 disciplines from the Physical Sciences and one from Life Sciences. The ratio of Physical Sciences to Life Sciences is 3:1, both in Profile 1 and Profile 2. The difference between the two profiles is that Profile 1 focuses on Physics, while in Profile 2, Computer Sciences is among the *Top disciplines*. The dominating disciplines of Profile 3 are Materials Sciences, Chemistry, and Biomedical Sciences, resulting in a 2:1 ratio of

Table 10 The mean proportions of disciplines and major fields (14 disciplines and 3 major fields) over profiles in %

	172 SCs	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5	Profile 6
14 Disciplines	Biomedical Sciences	***10***	***13***	***16***	***19***	***22***	***22***
	Clinical Sciences	3	7	8	***11***	***16***	***16***
	Neuro-Sciences	3	3	3	5	8	9
	Infectious Diseases	2	2	4	6	6	7
	Gen. Medicine; Health	1	3	3	2	4	6
	Materials Sciences	***21***	***19***	***15***	***10***	7	5
	Chemistry	***19***	***15***	***17***	***13***	8	7
	Computer Sciences	5	16	9	6	7	5
	Engineering	9	6	7	5	4	3
	Physics	***17***	6	6	7	5	4
	Ecology	3	1	1	5	3	5
	Environmental Sciences	1	4	4	4	4	3
	Geo-sciences	7	2	2	3	3	3
	Agriculture	1	2	4	5	3	4
3 Major fields	Life Sciences	*18*	*29*	*34*	*43*	*56*	*60*
	Physical Sciences	*71*	*62*	*54*	*41*	*31*	*24*
	Environmental Sciences	*11*	*9*	*12*	*16*	*13*	*16*

Physical Sciences to Life Sciences. The four *Top disciplines* of Profile 4 are Materials Sciences, Chemistry, Biomedical Sciences, and Clinical Sciences, the former two belonging to the Physical Sciences, and the latter two to the Life Sciences, resulting in a 1:1 ratio. This profile indeed leads to the most balanced situation. In Profile 5 and 6, the dominating disciplines are Biomedical Sciences, and Clinical Sciences, two disciplines from the Life Sciences.

We further note that South Korea has the most balanced disciplinary structure among the Asian countries in our study. Glänzel et al. (2008) analyzed the national subject profiles in the context of global changes in scientific research. These authors noted that South Korea and Taiwan tend to follow the Japanese profile, with a certain shift towards the profile of Western countries. The results of Glänzel et al. (2008) cannot be confirmed by our analysis, despite the two countries' concentration on Clinical Sciences. We think that our approach can identify the difference between national disciplinary structures on a smaller scale and hence reveals more information. In our study Japan is the only Asian country that is situated in a group of Western countries, showing that its disciplinary structure is still different from that of the other Asian countries. Interestingly, the disciplinary structures of Brazil, Russia, India, and China, forming the BRIC countries, do not belong to the same profile. Russia on its own forms Profile 1, China belongs to Profile 2 and India belongs to Profile 3, which has a close relationship with Profile 4 (including Japan). Brazil is an exception among the BRIC countries, due to its heavy focus on Life Sciences. This places it in Profile 6, together with the USA, the UK, Denmark, Australia, Norway and the Netherlands.

In Appendix B we provide a closer look at each of the six profiles separately.

Diversity of the disciplinary structure within countries and the relation with the profile to which they belong

Table 11 shows the values of the Simpson and Stirling–Rao diversity values of the 32 countries/regions and their rank. Clearly for each country the value of the Stirling–Rao diversity is always smaller than that of the Simpson diversity, confirming inequality (7). The Pearson correlation between the two diversity indices is 0.77. As the Simpson Index and the Cosine Index assume that the similarity between different categories is zero these measures overestimate the diversity of national disciplinary structures. We recall this was the reason why we prefer using the Stirling–Rao diversity measure to compare the balance in disciplinary structure of countries and profiles.

The higher the value of the diversity, the more balanced the countries' disciplinary structure. Considering ranks according to mean values per profile we notice that Simpson's measure follows profile numbers, which means that it follows the division based on preferences for Life Sciences, on the one hand, and Physical Sciences on the other. This observation does not hold for the Stirling–Rao measure. Considering specific countries we see that ranks according to Simpson's measure begin with Spain and Canada (highest diversity) and end with Russia, China and Singapore. According to the Stirling–Rao measure the most diverse countries are Turkey and Norway. Russia, China and Singapore again have the lowest diversity. The largest changes in rank occur for Turkey and Poland in one direction, and for the UK and the USA in the other.

We note that the values in Table 11 are all very close to 1. This is one of the drawbacks of using Simpson's measure. The larger the number of categories, the higher the Simpson diversity value (and the lower its concentration value). Using Stirling–Rao diversity only partially removes this drawback. This leads us to the open problem of finding another standard diversity measure which can be adapted to include disparity. We further note that because values are at times very close to one another we found that it makes a difference if diversity indices are calculated with four decimals or with higher precision. The results shown in Table 14 are based on calculations with 18 decimals, rounded to show four decimals.

Conclusions

The main contribution of this article is to provide a simple and easy-to-generalize analytical framework to study systems' balance and the homogeneity of systems.

Concretely we:

- Combine the work of Stirling (2007) with the models proposed by Nei and Li (1979) and Shriver et al. (1995), based on genetic distance.
- Propose a conceptual framework to reveal the relationships of systems by investigating the diversity within and among systems and the similarity between systems, based on the proportions of categories weighted by the similarity of categories. The field-similarity weighted cosine measure can take category dependencies into account.
- Perform an informetric study of national disciplinary structures.

As explained before, we have four goals in mind when performing the case study. We hope that we reach our first goal, namely providing a practical illustration of the use of our framework. Six profiles of national disciplinary structures among 32 countries/regions are identified. All profiles are characterized by disciplinary balance and prioritized disciplines, and they are

Table 11 Simpson and Stirling–Rao diversity values for 32 countries/regions

Country	Profile	Simpson diversity	Rank	Stirling–Rao diversity	Rank
Turkey	Profile 5	0.9874	16	0.8660	1
Norway	Profile 6	0.9880	12	0.8624	2
Spain	Profile 4	0.9892	1	0.8594	3
Finland	Profile 5	0.9889	3	0.8594	4
Portugal	Profile 4	0.9882	9	0.8570	5
Mexico	Profile 4	0.9879	13	0.8570	6
France	Profile 4	0.9886	6	0.8570	7
Poland	Profile 4	0.9861	23	0.8567	8
Canada	Profile 5	0.9891	2	0.8552	9
Iran	Profile 3	0.9860	25	0.8537	10
Australia	Profile 6	0.9888	4	0.8536	11
Greece	Profile 5	0.9872	17	0.8528	12
Belgium	Profile 5	0.9887	5	0.8508	13
Israel	Profile 5	0.9870	19	0.8483	14
Italy	Profile 5	0.9868	21	0.8462	15
Austria	Profile 5	0.9876	14	0.8452	16
Switzerland	Profile 5	0.9871	18	0.8445	17
Germany	Profile 5	0.9875	15	0.8441	18
Brazil	Profile 6	0.9883	8	0.8438	19
UK	Profile 6	0.9884	7	0.8428	20
Taiwan	Profile 2	0.9824	29	0.8419	21
Sweden	Profile 5	0.9880	10	0.8412	22
Czech Republic	Profile 4	0.9861	24	0.8395	23
Japan	Profile 4	0.9848	26	0.8356	24
USA	Profile 6	0.9880	11	0.8346	25
Denmark	Profile 6	0.9870	20	0.8339	26
South Korea	Profile 3	0.9829	28	0.8334	27
The Netherlands	Profile 6	0.9867	22	0.8331	28
India	Profile 3	0.9847	27	0.8295	29
Russia	Profile 1	0.9796	31	0.8179	30
China	Profile 2	0.9820	30	0.8176	31
Singapore	Profile 2	0.9783	32	0.8149	32
All countries in Profile 1	Mean value:	0.9796	1	0.8179	1
All countries in Profile 2	Mean value:	0.9809	2	0.8248	2
All countries in Profile 3	Mean value:	0.9845	3	0.8389	3
All countries in Profile 4	Mean value:	0.9873	4	0.8517	6
All countries in Profile 5	Mean value:	0.9878	5	0.8503	5
All countries in Profile 6	Mean value:	0.9879	6	0.8434	4
	Mean over all countries	0.9865		0.8446	

distributed in MDS mapping regularly. As to the second goal: there are indeed observed differences (for example Greece and Turkey) between the results obtained by the cosine similarity and the similarity-weighted cosine measure. In keeping with the third goal,

differences are explained by referring to 14 disciplines, the results of factor analysis from an independent study—Leydesdorff and Rafols (2009) who were not interested in relations between countries—but based on the same 172 SCs, validating our similarity results. The similarity-weighted cosine measure and factor analysis are two different ways to take into account category similarity. Considering the fourth goal, the diversity results using Simpson's measure or the Stirling–Rao measure are not very convincing as their differences are very small.

In addition we note that the BRIC countries are dispersed over different patterns. Russia, China, India and Brazil belong to Profile 1, Profile 2, Profile 3 and Profile 6, respectively. This shows that the term of BRIC countries is just a handy short-hand for a group of upcoming countries. It has, however, no structural meaning.

Following Burrell (2007) a next step would be to perform time dependent studies within the proposed framework. We hope that our approach can reach its full potential as a general conceptual framework for studying multidisciplinary research and its changes over time.

Acknowledgments This research was supported by the Young Talent Projects of the National Science Library, Chinese Academy of Sciences (No. YP08Z91001). The authors would like to thank Loet Leydesdorff, Andy Stirling, Ismael Rafols, and the anonymous reviewers for their helpful comments and Sven Romberg for modifying the syntax errors on an earlier version of the manuscript. Useful discussions with Jielan Ding are gratefully acknowledged. Data for the basic S_{ij} matrix were kindly provided by Loet Leydesdorff. Guido Herweyers provided the counterexample of a special symmetric matrix which is not positive definite. Rousseau's work is supported by NSFC grant No. 71173154. He is thankful for the hospitality provided to him during his research stay at the National Science Library, Chinese Academy of Sciences.

Appendix A

In this appendix we provide further necessary mathematical explanations.

Proposition 1 Using the notation introduced in the main text we show that: $\sum_{i,j=1}^{N} p_{X_i} p_{X_j} D_{ij} = 1 - \sum_{i,j=1}^{N} p_{X_i} p_{X_j} S_{ij}$

Proof $\sum_{i,j=1}^{N} p_{X_i} p_{X_j} D_{ij} = \sum_{i,j=1}^{N} p_{X_i} p_{X_j} (1 - S_{ij}) = \sum_{i=1}^{N} p_{X_i} \left(\sum_{j=1}^{N} p_{X_j} - \sum_{j=1}^{N} p_{X_j} S_{ij} \right)$
$= \sum_{i=1}^{N} p_{X_i} \left(1 - \sum_{j=1}^{N} p_{X_i} S_{ij} \right) = \sum_{i=1}^{N} p_{X_i} - \sum_{i,j=1}^{N} p_{X_i} p_{X_j} S_{ij} = 1 - \sum_{i,j=1}^{N} p_{X_i} p_{X_j} S_{ij}$

Proposition 2 Using the notation introduced in the main text we show that $L_X = 1 - \sum_{i=1}^{N} p_{X_i}^2$ can be derived from $SR_X = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} p_{X_i} p_{X_j} D_{ij}$ by taking all $D_{ij} = 1$ $(i \neq j)$.

Proof As $\sum_{i=1}^{N} p_{X_i} = 1$ we have: $1 = \left(\sum_{i=1}^{N} p_{X_i} \right)^2 = \sum_{i=1}^{N} p_{X_i}^2 + \sum_{i \neq j} p_{X_i} p_{X_j}$. Hence $1 - \sum_{i=1}^{N} p_{X_i}^2 = \sum_{i \neq j} p_{X_i} p_{X_j}$. The left-hand side of this equality is L_X, the Simpson diversity measure, while the right-hand side is SR_X in the case that for $i \neq j$ all $D_{ij} = 1$.

We would like to show that the expression used in Eq. 11, namely $\frac{\sum_{i,j=1}^{N} p_{X_i} p_{Y_j} S_{ij}}{\sqrt{\left(\sum_{i,j=1}^{N} p_{X_i} p_{X_j} S_{ij} \right) \cdot \left(\sum_{i,j=1}^{N} p_{Y_i} p_{Y_j} S_{ij} \right)}}$ is always smaller than or equal to 1. For this we recall the notion of a (real) inner product on a real vector space.

Definition If V is a real vector space then $\langle ., . \rangle : V \times V \rightarrow \mathbb{R}$ is a real inner product if

(1) $\forall x \in V : \langle x, x \rangle \geq 0$ with equality only if x = 0
(2) $\forall x, y \in V : \langle x, y \rangle = y, x$

(3) $\forall x, y, z \in V$ and $c \in \mathbb{R} : \langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ and $\langle cx, y\rangle = c\langle x, y\rangle$

If x, y belong to $\mathbb{R}^n$ then $x, y = \sum_{i=1}^{n} x_{i.} y_i$, where $(x_i)_i$ and $(y_i)_i$ are the components of x and y, is the standard inner product in $\mathbb{R}^n$. In an inner product space (a vector space equipped with an inner product) one can define a norm, denoted as $||\,.\,||$, by $||x|| = \sqrt{\langle x, x\rangle}$. In such spaces the Cauchy–Schwarz–Bunyakowski (C–S–B) inequality is valid: $||\langle x, y\rangle|| \leq ||x||.||y||$. Rewriting C–S–B using the standard inner product in $\mathbb{R}^n$, yields: $\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_{i.}^2}\sqrt{\sum_{i=1}^{n} y_{i.}^2}} \leq 1$ if all components of x and y are non-negative (and x and y are not both equal to the zero vector). This inequality just shows that the cosine measure always gives values between -1 and $+1$; and in the case of vectors with non-negative components yields values between 0 and 1.

It can be shown that if M is an N × N symmetric, positive definite matrix, then the expression $\langle x|y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{ij} y_j$ defines an inner product in $\mathbb{R}^n$ (of course in general different from the standard one). For symmetric matrices being positive definite is equivalent to having only strictly positive eigenvalues. The matrix S expressing the similarity between SCs is symmetric, and has moreover the following two properties: its diagonal contains only the value 1; and the values outside the diagonal are all numbers between 0 (included) and 1 (not included). Is such a matrix positive definite? Unfortunately the answer is no. Already in three dimensions it is possible to give counterexamples. The following matrix is symmetric and satisfies the extra requirements:

$$M_{\mathrm{H}} = \begin{pmatrix} 1 & 0.8 & 0.9 \\ 0.8 & 1 & 0 \\ 0.9 & 0 & 1 \end{pmatrix}$$

Yet, the eigenvalues of M_{H} are: 2.20416; 1 and -0.204159. This means that it is not possible to use an arbitrary matrix expressing the similarity between categories to form an inner product. Fortunately, the specific S-matrix we used IS positive definite (using Matlab we checked that all its eigenvalues are strictly positive). Hence we may define an inner product satisfying the Cauchy–Schwarz–Bunyakowski inequality:

$$\left| \frac{\sum_{i,j=1}^{N} x_i y_j S_{ij}}{\sqrt{\left(\sum_{i,j=1}^{N} x_i x_j S_{ij}\right).\left(\sum_{i,j=1}^{N} y_i y_j S_{ij}\right)}} \right| \leq 1$$

This implies that values of the weighted cosine measure, as used in this contribution, are always situated between 0 and 1.

Next we would like to show that standard cosine values are never larger than the corresponding values for the weighted cosine measure. This turns out to be true for all cases that occurred here, so that we can indeed say that the standard cosine measure underestimates the similarity between countries. Yet we are not able to show this mathematically (for this particular S). We leave this as an open problem.

Appendix B: closer look at each of the six profiles separately

To further explain the map of the six profiles, the proportions of 32 countries/regions in 14 disciplines and 3 major fields are shown in Table 12, 13, 14, and 15. *Top disciplines* of each country have been put in bold italics.

For Profile 2, which includes China, Singapore and Taiwan, the four *Top disciplines* of the three countries/regions are the same, but the ranks of the four *Top disciplines* are not the same. Actually, among these three China's disciplinary structure is the most similar to that of Russia (Profile 1). China has Materials Sciences and Chemistry both much more prioritized than the other two. Materials Sciences is the dominant discipline of Singapore, and Computer Science ranks second. The disciplinary structure of Taiwan is very similar to Singapore's, the only difference being Computer Sciences' rank as its number one *Top discipline*.

The third profile group, including India, Iran and South Korea, exhibits more evenness than Profiles 1 and 2. Iran has Chemistry as the number one *Top discipline*, but the proportions of Computer Science, Materials Sciences and Biomedical Sciences are almost equal. India has Chemistry as the number one *Top discipline*, but with more focus on Biomedical Sciences. South Korea pays the most attention to Materials Sciences and Biomedical Sciences. Clinical Sciences is the fourth *Top discipline* of this country.

Profile 4 is a cluster including Japan, the Czech Republic, Spain, France, Poland, Mexico and Portugal, which can be divided into two parts according to the 3 major fields (see Table 13). One part, Sect. 1, consists of Japan, the Czech Republic, Spain and France. They pay more attention to the Life Sciences with a mean proportion of Life Sciences in the four countries at 47 % ± 2.4, and a mean proportion of Physical Sciences in the four

Table 12 The proportions of countries/regions in Profiles 1–3 in 14 disciplines and 3 major fields (in %)

	172 SCs	Profile 1		Profile 2		Profile 3		
		Russia	China	Singapore	Taiwan	Iran	South Korea	India
4 Disciplines	Biomedical Sciences	***10***	***13***	***15***	***12***	***11***	***18***	***18***
	Clinical Sciences	3	4	9	9	7	***11***	7
	Neuro-Sciences	3	2	4	4	3	5	2
	Infectious Diseases	2	2	2	3	6	3	4
	Gen. Medicine; Health	1	1	5	3	4	2	2
	Materials Sciences	***21***	***19***	***21***	***18***	***12***	***19***	***15***
	Chemistry	***19***	***21***	***13***	***9***	***18***	***13***	***19***
	Computer Sciences	5	***10***	***16***	***21***	***11***	***11***	5
	Engineering	9	6	5	7	11	5	5
	Physics	***17***	9	5	5	6	6	7
	Ecology	3	2	1	1	1	1	1
	Environmental Sciences	1	4	3	4	5	3	4
	Geo-sciences	7	3	1	2	2	1	3
	Agriculture	1	3	1	2	4	3	6
3 Major fields	Life Sciences	18	23	34	30	30	39	34
	Physical Sciences	71	66	61	60	58	53	51
	Environmental Sciences	11	11	5	10	12	8	14

Table 13 Proportions of countries in Profile 4 in 14 disciplines and 3 major fields(in %)

	172 SCs	Poland	Mexico	Portugal	Japan	Czech Republic	France	Spain
14 Disciplines	Biomedical Sciences	***17***	***16***	***17***	***22***	***23***	***21***	***18***
	Clinical Sciences	***10***	5	***9***	***15***	***9***	***14***	***12***
	Neuro-Sciences	5	4	4	7	5	6	6
	Infectious Diseases	5	7	6	4	5	6	7
	Gen. Medicine; Health	2	4	2	1	1	3	4
	Materials Sciences	***10***	***12***	***10***	***13***	***11***	***9***	8
	Chemistry	***16***	***11***	***14***	***12***	***14***	***9***	***11***
	Computer Sciences	8	5	6	6	5	7	8
	Engineering	6	4	6	3	4	5	4
	Physics	8	8	5	6	6	7	5
	Ecology	3	***9***	6	3	4	3	4
	Environmental Sciences	4	4	6	2	3	3	4
	Geo-sciences	2	3	4	2	3	4	3
	Agriculture	4	7	4	3	6	3	5
3 Major fields	Life Sciences	39	36	39	49	44	49	47
	Physical Sciences	47	42	41	41	39	38	36
	Environmental Science	14	23	20	10	17	12	17

countries at 39 % ± 1.2. The other section of Profile 4 is Poland, Mexico and Portugal, with a mean proportion of Life Sciences in the four countries at 38 % ± 1.4, and a mean proportion of Physical Sciences in the four countries at 43 % ± 2.6. So the disciplinary structure of the second part, Sect. 2, is more balanced than Sect. 1. The countries in Profile 4 all have Biomedical Sciences as their prime discipline, but the countries from Sect. 2 are more focused on Chemistry than those belonging to Sect. 1.

As mentioned earlier we notice that South Korea has the most balanced disciplinary structure among the Asian countries in our study. Japan's too belongs to profile 4. Its disciplinary structure is more like Spain's or France's because they all have Clinical Sciences as their second *Top discipline*.

The 11 countries constituting Profile 5 are displayed in Table 12. For each country, the proportion of Biomedical Sciences and Clinical Sciences is even larger than 10 % and the proportion of Biomedical Sciences is larger than 20 %. Profile 5 has more polarized countries than Profile 4, with Biomedical Sciences and Clinical Sciences as dominant disciplines.

One notable feature of Profile 6 is its inclusion of Brazil. Brazil's growing economy places it within Profile 6. It is similar to the USA, the UK, Denmark, Australia, Norway and the Netherlands, where Biomedical Sciences and Clinical Sciences are commonly the focus of activity, a typical characteristic of Profile 6. Yet, Brazil is still located in the margin of the MDS map. Agriculture is the fourth dominant discipline of Brazil, which is different from the other countries in Profile 6.

Within this profile the UK is very similar to the USA. Along with Brazil, Norway is another country showing different characteristics from the other countries in Profile 6 as ecology is its third priority discipline.

Table 14 The proportions of countries in Profile 5 in 14 disciplines and 3 major fields (in %)

	172 SCs	Turkey	Greece	Israel	Finland	Germany	Italy	Belgium	Austria	Canada	Switzerland	Sweden
14 Disciplines	Biomedical Sciences	***17***	***23***	***23***	***20***	***22***	***22***	***22***	***22***	***21***	***20***	***25***
	Clinical Sciences	***19***	***18***	***14***	***12***	***17***	***19***	***16***	***19***	***15***	***14***	***14***
	Neuro-Sciences	6	5	***10***	8	9	8	7	8	9	***10***	8
	Infectious Diseases	6	5	5	5	5	5	7	6	4	6	6
	Gen. Medicine; Health	4	3	4	4	3	2	4	4	6	7	5
	Materials Sciences	7	8	7	8	***9***	7	7	7	6	7	7
	Chemistry	***10***	8	7	8	***10***	8	8	8	7	***9***	7
	Computer Sciences	8	***10***	8	8	4	7	7	6	8	4	5
	Engineering	6	4	5	3	3	4	3	3	4	3	4
	Physics	4	4	7	6	7	6	5	5	4	5	5
	Ecology	2	2	2	6	2	2	4	3	5	3	5
	Environmental Sciences	6	5	2	4	2	3	3	3	4	3	3
	Geo-sciences	2	3	2	3	3	3	2	3	4	5	3
	Agriculture	4	3	3	4	3	3	4	3	4	3	3
3 Major fields	Life Sciences	52	53	56	50	57	57	56	59	55	58	59
	Physical Sciences	35	34	34	33	33	32	31	29	28	28	28
	Environmental Sciences	14	13	10	18	10	11	13	12	17	14	14

Table 15 Proportions of countries in Profile 6 in 14 disciplines and 3 major fields (in %)

	172 SCs	Brazil	UK	USA	Australia	Denmark	Norway	The Netherlands
14 Disciplines	Biomedical Sciences	***21***	***23***	***25***	***19***	***26***	***18***	***23***
	Clinical Sciences	***13***	***17***	***17***	***14***	***15***	***14***	***23***
	Neuro-Sciences	7	***10***	9	***11***	8	***10***	***11***
	Infectious Diseases	***10***	6	5	6	8	6	6
	Gen. Medicine; Health	4	7	6	8	5	5	5
	Materials Sciences	6	6	6	6	6	4	5
	Chemistry	***9***	6	6	6	7	6	6
	Computer Sciences	4	5	5	5	4	7	5
	Engineering	3	4	3	3	3	3	3
	Physics	4	5	4	3	4	3	3
	Ecology	6	4	4	7	5	***10***	3
	Environmental Sciences	3	2	3	4	4	5	3
	Geo-sciences	2	3	3	4	2	6	2
	Agriculture	9	2	3	4	4	3	2
3 Major fields	Life Sciences	55	63	63	58	62	53	68
	Physical Sciences	26	26	25	24	24	22	22
	Environmental Sciences	19	11	13	19	15	24	10

The common features of Profiles 5 and 6 can be found when comparing Tables 14 and 15. These two profiles have Life Sciences as their most dominant major fields, while their main difference lies in the fact that the disciplinary structure of Profile 5 is more balanced than Profile 6.

References

Ackoff, R. L. (1960). Systems, organizations, and interdisciplinary research. *Society for General Systems Research, 5*, 1–8.

Almeida, J. A. S., Pais, A. A. C. C., & Formosinho, S. J. (2009). Science indicators and science patterns in Europe. *Journal of Informetrics, 3*(2), 134–142.

Barjak, F., & Robinson, S. (2008). International collaboration, mobility and team diversity in the life sciences: Impact on research performance. *Social Geography, 3*, 23–36.

Barjak, F., & Robinson, S. (2009). National cultural diversity of research teams. In B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009—12th international conference of the international society for scientometrics and informetrics* (pp. 381–393). São Paulo: BIREME/PAHO/WHO & Federal University of Rio de Janeiro.

Bordons, M., Morillo, F., & Gómez, I. (2004). Analysis of cross-disciplinary research through bibliometric tools. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 437–456). Dordrecht: Kluwer.

Burrell, Q. L. (2005). Measuring informetric similarity of concentration between different distributions: Two new approaches. *Journal of the American Society for Information Science, 56*(7), 704–714.

Burrell, Q. L. (2006). Measuring concentration within and co-concentration between informetric distributions: An empirical study. *Scientometrics, 68*(3), 441–456.

Burrell, Q. L. (2007). Time-dependent aspects of co-concentration in informetrics. *Scientometrics, 73*(2), 161–174.
Clarke, A. (1978). *Analytical archaeology*. London: Methuen.
Dagum, C. (1987). Measuring the economic affluence between populations of income receivers. *Journal of Business and Economic Statistics, 5*(1), 5–11.
Dalton, H. (1920). The measurement of inequality of incomes. *The Economic Journal, 30*, 348–361.
Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science, 42*(7), 479–489.
Finkelstein, M., & Friedman, R. (1967). The application of an entropy theory of concentration to the Clayton Act. *Yale Law Journal, 76*, 677–717.
Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia, 2*, 143–148.
Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69–115.
Glänzel, W., Debackere, K., & Meyer, M. (2008). 'Triad' or 'tetrad'? On global changes in a dynamic world. *Scientometrics, 74*(1), 71–88.
Gregorius, H.-R. (1987). The relationship between the concepts of genetic diversity and differentiation. *Theoretical and Applied Genetics, 74*(3), 397–401.
Griffiths, A., Robinson, L. A., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*(3), 175–205.
Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology, 54*, 427–432.
Hu, X. J., & Rousseau, R. (2009). Comparative study of the difference in research performance in biomedical fields among selected Western and Asian countries. *Scientometrics, 81*(2), 475–491.
King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6), 311–316.
Kozlowski, J., Radosevic, S., & Ircha, D. (1999). History matters: the inherited disciplinary structure of the post-communist science in countries of Central and Eastern Europe and its restructuring. *Scientometrics, 45*(1), 137–166.
Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos, 76*(1), 5–13.
Leonard, R. D., & Jones, G. T. (Eds.). (1989). *Quantifying diversity in archaeology*. Cambridge: Cambridge University Press.
Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology, 58*(9), 1303–1319.
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.
Liang, L. M., Havemann, F., Heinz, M., & Wagner-Döbler, R. (2006). Structural similarities between science growth dynamics in China and in western countries. *Scientometrics, 66*(2), 311–325.
Liu, Yx., Rafols, I., & Rousseau, R. (2012). A framework for knowledge integration and diffusion. *Journal of Documentation, 68*(1), 31–44.
Magurran, A. (1988). *Ecological diversity and its measurement*. London: Croom Helm.
McIntosh, R. (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology, 48*, 392–404.
Morillo, F., Bordons, M., & Gomez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics, 51*(1), 203–222.
Morillo, F., Bordons, M., & Gomez, I. (2003). Interdisciplinarity in science: a tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology, 54*(13), 1237–1249.
Nei, M. (1972). Genetic distance between populations. *The American Naturalist, 106*(949), 283–292.
Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, 70*, 3321–3323.
Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics, 89*, 583–590.
Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, 76*(10), 5269–5273.
Nei, M., & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics, 97*, 145–163.
Nguyen, P., Saviotti, P. P., Trommetter, M., & Bourgeois, B. (2005). Variety and the evolution of refinery processing. *Industrial and Corporate Change, 14*(3), 469–500.
Nijssen, D., Rousseau, R., & Van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses, 13*(1), 33–38.

Patil, G., & Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association, 77*(379), 548–567.
Pielou, E. (1969). *An introduction to mathematical ecology*. New York: Wiley.
Porter, A. L., & Chubin, D. E. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*, 161–176.
Porter, A. L., Cohen, A. S., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics, 72*(1), 117–147.
Qin, J., Lancaster, F. W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for Information Science, 48*(10), 893–916.
Rafols, I., & Meyer, M. (2007). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics, 70*(3), 633–650.
Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.
Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology, 21*(1), 24–43.
Rousseau, R. (1992). Concentration and diversity of availability and use in information systems—a positive reinforcement model. *Journal of the American Society for Information Science, 43*(5), 391–395.
Rousseau, R., & Van Hecke, P. (1999). Measuring diversity. *Acta Biotheoretica, 47*, 1–5.
Rousseau, R., Van Hecke, P., Nijssen, D., & Bogaert, J. (1999). The relationship between diversity profiles, evenness and species richness based on partial ordering. *Environmental and Ecological Statistics, 6*(2), 211–223.
Runnegar, B. (1987). Rates and modes of evolution in the mollusca. In K. S. W. Campbell & M. F. Day (Eds.), *Rates of evolution* (pp. 39–60). London: Allen and Unwin.
Segev, E., Ahituv, N., & Barzilai-Nahon, K. (2007). Mapping diversities and tracing trends of cultural homogeneity/heterogeneity in cyberspace. *Journal of Computer-mediated communication, 12*(4), 1269–1297.
Shriver, M. D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R. E., & Chakraborty, R. (1995). A novel measure of genetic distance for highly polymorphic tandem. *Molecular Biology and Evolution, 12*(5), 914–920.
Simpson, E. H. (1949). Measurement of diversity. *Nature, 688*, 163.
Sokal, R., & Sneath, P. (1970). *Principles of numerical taxonomy* (pp. 20–22). San Francisco: Freeman.
Solow, A., & Polasky, S. (1994). Measuring biological diversity. *Environmental and Ecological Statistics, 1*, 95–107.
Stanley, M. H. R., Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., et al. (1996). Scaling behaviour in the growth of companies. *Nature, 379*(6568), 804–806.
Stirling, A. (1998). On the economics and analysis of diversity. SPRU Electronic Working Paper. http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp28/sewp28.pdf. Accessed 4 Sep 2011.
Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719.
van Leeuwen, T., & Tijssen, R. (2000). Interdisciplinary dynamics of modern science: Analysis of cross-disciplinary citation flows. *Research Evaluation, 9*(3), 183–187.
van Raan, A. F. J., & van Leeuwen, T. (2002). Assessment of the scientific basis of interdisciplinarity, applied research. Application of bibliometric methods in Nutrition and Food Research. *Research Policy, 31*, 611–632.
Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs, 30*, 279–338.
Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon, 21*, 213–251.
Williams, P. H., & Humphries, C. J. (1994). Biodiversity, taxonomic relatedness and endemism in conservation. In L. Forey, C. Humphries, & R. Vane-Wright (Eds.), *Setting objectives for conservation evaluation* (pp. 269–287). Oxford: Oxford University Press.
Yue, H. J. (2008). Prediction and relative dynamic performance on field of research of China. *Studies in Science of Science, 26*(3), 530–538.
Yue, T., Yang, L. Y., Ding, J. L., & Han, T. (2011). Comparison of disciplinary structure in science between G7 and the BRIC countries by bibliometric methods. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2011 Conference* (pp. 887–897). Durban: ISSI, Leiden University & University of Zululand.
Zhang, L., Rousseau, R., & Glänzel, W. (2011). Document-type country profiles. *Journal of the American Society for Information Science and Technology, 62*(7), 1403–1411.
Zitt, M. (2005). Facing diversity of science: A challenge for bibliometric indicators. *Measurement, 3*(1), 38–49.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*, 373–401.

Zitt, M., Bassecoulard, E., Filliatreau, G., & Ramanana-Rahary, S. (2007). Revisiting country and institution indicators from citation distributions: Profile performance measures. In D. Torres-Salinas & H. F. Moed (Eds.), *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics* (ISSI 2007) (pp. 797–802). Madrid, Spain: Centre for Scientific Information and Documentation, Spanish Research Council.