Scientometrics manuscript No.
(will be inserted by the editor)

Benchmarking Research Performance at the University Level with Information Theoretic Measures

J. A. García, Rosa Rodriguez-Sánchez,

J. Fdez-Valdivia, Nicolas Robinson-García,

and Daniel Torres-Salinas.

the date of receipt and acceptance should be inserted later

Abstract This paper presents a new method for comparing universities based on information theoretic measures. The research output of each academic institution is represented statistically by an impact-factor histogram. To this aim, for each academic institution we compute the probability of occurrence of a publication with impact factor in different intervals. Assuming the probabilities associated with a pair of academic institutions our objective is to measure the Information

J. A. García, Rosa Rodriguez-Sánchez, J. Fdez-Valdivia,

Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, 18071 Granada, Spain.

Address correspondence to J. A. García at jags@decsai.ugr.es

Nicolas Robinson-García

EC3: Evaluación de la Ciencia y la Comunicación Científica,
Universidad de Granada, 18071 Granada, Spain

Daniel Torres-Salinas

EC3: Evaluación de la Ciencia y la Comunicación Científica, Centro de Investigación Médica Aplicada, Universidad de Navarra, 31008, Pamplona, Navarra, Spain

Gain between them. To do so, we develop an axiomatic characterization of relative information for predicting institution-institution dissimilarity. We use the Spanish university system as our scenario to test the proposed methodology for benchmarking three universities with the rest as a case study. For each case we use different scientific fields such as Information and Communication Technologies, Medicine & Pharmacy, and Economics & Business as we believe comparisons must take into account their disciplinary context. Finally we validate the Information Gain values obtained for each case with previous studies.

Keywords Information gain, institution-institution similarity, impact-factor histogram, information theoretic measure, information conservation constraint, benchmarking research output.

1 Introduction

The emergence of a highly competitive environment in which universities compete for top-talented students, researchers and research funding worldwide (Hazelkorn, 2011), along with the economic constraints countries are suffering, has lead to the development of numerous tools for benchmarking and monitoring research activities at an institutional level. Originally conceived as elite institutions for highly qualified education, universities have evolved into complex institutions that combine traditional roles inherited by their own history with the demands of current times, transforming them into flexible and dynamic entities capable of generating wealth to their surroundings. In this sense, the launch of national rankings in the late eighties and international rankings at the beginning of the 21st Century along with the development of the so-called "evaluative bibliometrics" (Lundberg,

2006) and the introduction of national research assessment exercises in different countries (Moed, 2008; Abramo et al., 2011; Vanclay and Bornmann, 2012), have centred the development of novel methodologies for research monitoring in the spotlight of the research policy agenda.

Many studies can be found in the literature developing such techniques in order to classify academic institutions (Shin, 2009; Ortega et al., 2011), establish institutional profiles (Carpenter et al., 1988; Garcia et al., 2012a) or compare universities performance (Adams et al., 2007; Tijssen et al., 2009; Torres-Salinas et al., 2011a). For instance, (Adams et al., 2007) present a methodology for profiling universities according to their impact as an intuitive way to identify outstanding institutions when compared with the average research performance of a country, region, or any other given research unit. (Turner, 2007) suggests using data envelope analysis (DEA) for benchmarking universities. In this paper, we propose applying dissimilarity techniques to universities research output in order to detect those which look alike according to their competitiveness defined as their capacity for publishing in top journals. We follow the line of thought proposed by (Carpenter et al., 1988) in which they state that journals' impact factor is a good proxy measure for large aggregations of data for research excellence.

It often happens that the structure of research-output of certain academic institution cannot be accurately determined due to various reasons: some of the details may not be observable or the researcher who makes an attempt to investigate the structure of research-output may not take all the relevant factors governing the structure for certain academic institution into consideration. Under such circumstances, the structure of research-output of universities or other academic institutions can be characterized statistically by impact-factor histograms.

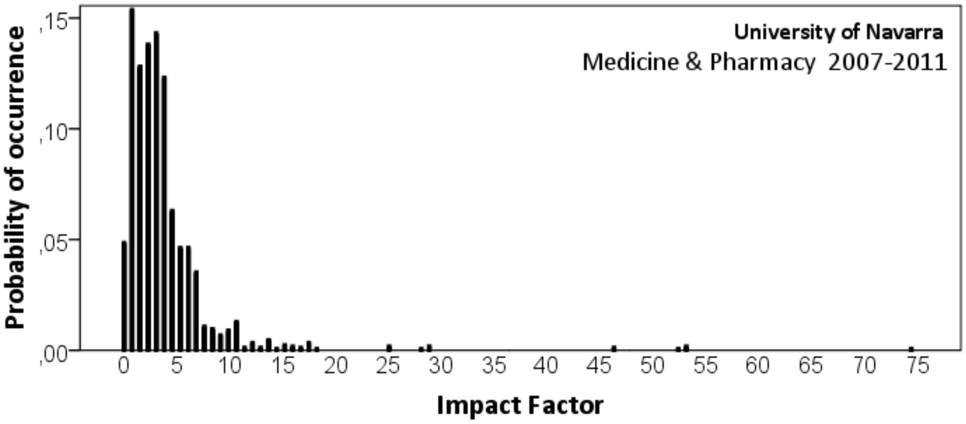

Fig. 1 Impact-factor histogram representing University of Navarra in the field of Medicine and Pharmacy

For instance we can compute the probability of occurrence of a publication with impact factor in the interval $[l_i, l_i + \Delta l)$, with $i = 0, 1, \cdots, n$, for each academic institution, and where $l_{i+1} = l_i + \Delta l$. Figure 1 shows the impact-factor histogram representing a Spanish university (Navarra) in Medicine & Pharmacy.

Let us assume the discrete probabilities associated with a reference university R and another university of input I as those given by P and Q, respectively (see Fig. 2). Here we ask the following question: What is the amount of relative information between P and Q? For this purpose a large number of measures has been developed by (Jeffreys, 1946; Kullback and Leibler, 1951; Renyi, 1961; Havrda and Charvat, 1967; Kapur, 1984; Sharma and Mittal, 1977; Burbea and Rao, 1983; Rao, 1982), and others. This makes it very difficult when choosing the criteria in order to see which one suits better. In order to do so, it is important to know which postulates and properties should be satisfied by the information theoretic measure. In this paper we attempt at developing an axiomatic characterization of

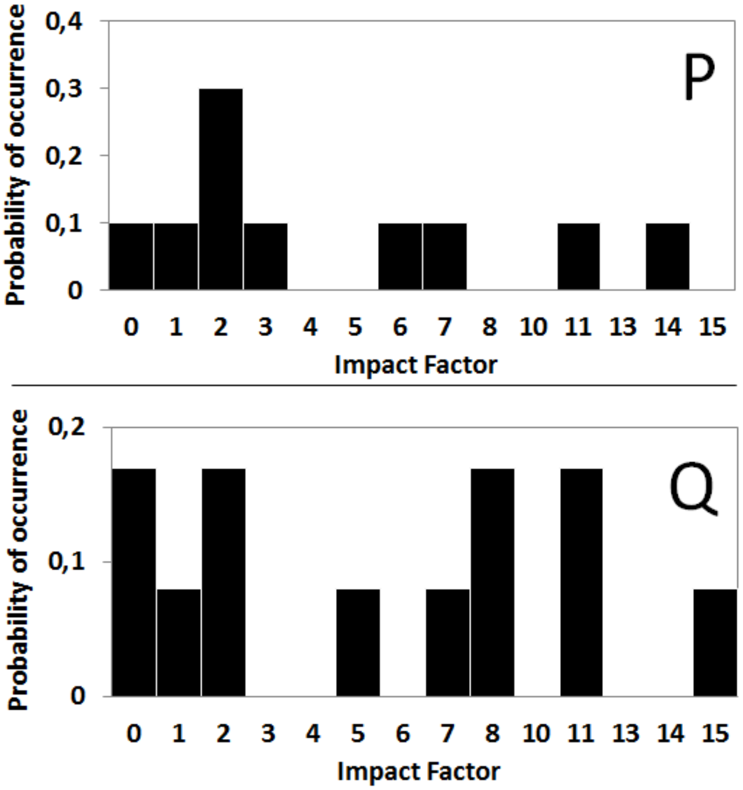

Fig. 2 $\mathcal{P} = \{p(l_i/R)|i = 0, 1, ..., 19\}$ and $\mathcal{Q} = \{p(l_i/I)|i = 0, 1, ..., 19\}$ which were computed using impact-factor intervals $[l_i, l_i + \triangle l)$, with $i = 0, 1, \ldots, 19$, as given in Table 1.

relative information for predicting institution-institution dissimilarities from their respective research outputs.

The paper is structured as follows. In Section 2 we describe the simplest form of information gain between two academic institutions, which is based on three basic postulates. A more advanced form of relative information (Section 3) is derived from the addition to the basic axiomatic characterization of a postulate relative to the effectiveness of the information as well as the information conservation constraint (i.e., the properties of an academic institution in significant fields of study are similar to the properties of another institution in the same fields). We also address the properties of the information measures and examine the implications of these properties. In each case, the minimum value of information gain between two institutions leads to the most similar academic institutions. In Section 4 we apply the information theoretic measures for benchmarking and comparing academic institutions focusing on three Spanish universities (Navarra, Granada and Pompeu Fabra) and three scientific fields (Medicine & Pharmacy, Information & Communication Technologies, and Economics & Business) respectively. In each case we show which universities are more alike to those which are employed as subject cases. We use a sample of 57 Spanish universities and the 2007-2011 study time period. Then, we validate this model by comparing our results with those from (Garcia et al., 2012a). Section 5 concludes with a discussion over the obtained results.

2 Basic Axiomatic Characterization For Measuring Information Gain Between Academic Institutions

This section presents the basic axiomatic characterization of a measure of information gain between an input academic institution I and another of reference R, where information gain measures the degree of dissimilarity between these two academic institutions. If we predict the similarity between academic institutions based on their information gain, then the minimum value of information gain between two institutions leads to the most similar academic institutions.

The objective is twofold: firstly, to characterize the information gain between two probability distributions (representing each one of the academic institutions as shown in Figure 1) with a minimal number of properties which are natural and thus desirable; and secondly, to determine the form of all error functions satisfying these properties which we have stated to be desirable for predicting institution-institution dissimilarity. This analysis is original to benchmark research performance and based on a formal approach for predicting visual target distinctness in Computer Vision (Garcia et al., 2001).

The first postulate states a property of how unexpected a single event of an academic institution was.

Axiom 1. *A measure $\mathcal{U}$ of how unexpected the single event "a publication with impact factor in the interval $[l, l + \Delta l)$ occurs" was, depends only upon its probability p.*

This means that there exists a function h defined in $[0, 1]$ such that

$$\mathcal{U}(\text{"a publication with impact factor in the range } [l, l + \Delta l) \text{ occurs"}) = h(p). \quad (1)$$

This is a natural property because we assume that the academic institutions are characterized by discrete probability distributions (e.g., impact-factor histograms) as shown in Fig. 1.

Our second postulate is formulated to obtain a reasonable estimate of how unexpected an academic institution was from some probability distribution by means of the mathematical expectation of how unexpected its single events were from this distribution.

Let $p(l/R)$ and $p(l/I)$ be the probability of occurrence of a publication with impact factor in the interval $[l, l + \Delta l)$ for a reference academic institution R and the input institution I, respectively. Fig. 2 shows two examples of impact-factor histograms. Suppose that every possible observation from $p(l/R)$ is also a possible observation from $p(l/I)$.

As stated in Axiom 1, if the single events of the reference institution R are characterized by an "estimated" distribution $\mathcal{Q} = \{p(l_i/I) \mid i = 0, 1, \cdots, n\}$, then the function $h(p(l_i/I))$, with $i = 0, 1, \cdots, n$, returns a measure of how unexpected each single event "a publication with impact factor in the interval $[l_i, l_i + \Delta l)$ occurs" was from $\mathcal{Q}$. Recall that $l_{i+1} = l_i + \Delta l$. Thus, assuming that $\mathcal{P} = \{p(l_i/R) \mid i = 0, 1, \cdots, n\}$ is the "true" probability distribution of the reference academic institution R, we have that:

Axiom 2. *An estimate of how unexpected the research performance of a reference academic institution was from some probability distribution is simply defined as the mathematical expectation of how unexpected its single events were from this distribution.*

That is, the mathematical expectation of the discrete random variable $h(Q)$, which can assume the values

$$h(p(l_0/I)), h(p(l_1/I)), \cdots, h(p(l_n/I))$$

with respective probabilities

$$p(l_0/R), p(l_1/R), \cdots, p(l_n/R)$$

is an estimate $\mathcal{U}_{\mathcal{P}}(\mathcal{Q})$ of how unexpected the reference academic institution R was from $\mathcal{Q} = \{p(l/I)\}$, i.e.,

$$\mathcal{U}_{\mathcal{P}}(\mathcal{Q}) = E_{\mathcal{P}}\left[h(\mathcal{Q})\right] = \sum_{l} p(l/R)\, h(p(l/I)) \tag{2}$$

with $E_{\mathcal{P}}$ denoting the mathematical expectation in $\mathcal{P}$.

The following postulate relates the estimate of how unexpected the reference academic institution was from an "estimated" distribution and the estimate from the "true" distribution.

Axiom 3. *The reference academic institution R with "true" probability distribution $\mathcal{P}$ is more unexpected from an "estimated" distribution $\mathcal{Q}$ than from the "true" distribution $\mathcal{P}$.*

The following inequality expresses how the reference academic institution is more unexpected when it is characterized by $\mathcal{Q}$ than when is characterized by $\mathcal{P}$:

$$\mathcal{U}_{\mathcal{P}}(\mathcal{Q}) \geq \mathcal{U}_{\mathcal{P}}(\mathcal{P}). \tag{3}$$

with $\mathcal{U}_{\mathcal{P}}(\mathcal{Q})$ and $\mathcal{U}_{\mathcal{P}}(\mathcal{P})$ being estimates of how unexpected the reference academic institution was from the "estimated" distribution $\mathcal{Q}$ and from the "true" distribution $\mathcal{P}$, respectively.

The true distribution $\mathcal{Q}$ of the input academic institution I may be interpreted as an estimated distribution of the reference institution R (with "true" distribution $\mathcal{P}$). Thus, we can define a measure of information gain of the reference academic institution from the input institution by the difference between the estimate of how unexpected the reference academic institution was from $\mathcal{Q}$ and that from $\mathcal{P}$.

Definition 1: A measure of information gain between academic institutions. Given the reference academic institution R with "true" probability distribution $\mathcal{P} = \{p(l/R)\}$, a measure of the information gain of the reference institution R from the input institution I with "true" distribution $\mathcal{Q} = \{p(l/I)\}$, is:

$$\mathcal{E}(\mathcal{P}, \mathcal{Q}) = \mathcal{U}_{\mathcal{P}}(\mathcal{Q}) - \mathcal{U}_{\mathcal{P}}(\mathcal{P}), \tag{4}$$

with $\mathcal{U}_{\mathcal{P}}(\mathcal{Q})$ and $\mathcal{U}_{\mathcal{P}}(\mathcal{P})$ being estimates of how unexpected the reference academic institution was from $\mathcal{Q}$ and $\mathcal{P}$, respectively. $\mathcal{U}_{\mathcal{P}}(\mathcal{Q})$ and $\mathcal{U}_{\mathcal{P}}(\mathcal{P})$ are defined as given in Axiom 2, and such that satisfy the inequality (3) in Axiom 3.

The following result serves to determine the form of the measure $\mathcal{E}(\mathcal{P}, \mathcal{Q})$.

Theorem 1. Let $\mathcal{E}(\mathcal{P},\mathcal{Q})$ be a measure of information gain for the discrimination between two academic institutions as given in Definition 1, i.e.,

$$\mathcal{E}(\mathcal{P},\mathcal{Q}) = \mathcal{U}_{\mathcal{P}}(\mathcal{Q}) - \mathcal{U}_{\mathcal{P}}(\mathcal{P}),$$

with $\mathcal{P} = \{p(l/R)\}$ and $\mathcal{Q} = \{p(l/I)\}$. Then, the measure of relative information $\mathcal{E}$ is equal to the Kullback-Leibler's information function (Kullback, 1978) between $\mathcal{P}$ and $\mathcal{Q}$ up to a nonnegative multiplicative constant, i.e.,

$$\mathcal{E}(\mathcal{P},\mathcal{Q}) = a\, E_{\mathcal{P}}\left[\log \frac{\mathcal{P}}{\mathcal{Q}}\right] \tag{5}$$

with $a \geq 0$ and $E_{\mathcal{P}}$ denoting the mathematical expectation.

Proof. See Theorem 1 in (Garcia et al., 2001)

In conclusion, any measure of how unexpected an academic institution was, that satisfies Axioms 1, 2, and 3, has to be of the form of the Kullback-Leibler information function up to a nonnegative multiplicative constant.

Table 1 illustrates an example of the computation of the information gain following Definition 1, for a pair of discrete probability distributions given in Fig. 2. That is, the structure of research-output of a pair of academic institutions of example is characterized statistically by discrete probability distributions, and Fig. 2 shows the probability of occurrence of a publication with impact factor in the interval $[l_i, l_i + \Delta l)$, with $i = 0, 1, \cdots, n$, for each academic institution, and where $l_{i+1} = l_i + \Delta l$.

In Fig. 2, $\mathcal{P} = \{p(l/R)\}$ and $\mathcal{Q} = \{p(l/I)\}$, where $p(l/R)$ and $p(l/I)$ denotes the probability of occurrence of a publication with impact factor in the interval $[l, l + \Delta l)$ for a reference academic institution R and the input institution I, respectively.

Table 1 presents the set of impact-factor intervals which were used in order to define the discrete probability distribution for each academic institution in this

$[l_i, l_i + \triangle l), i = 0, 1, \ldots, 19$ with $\triangle l = 0.75$	$p(l/R)log\left(\frac{p(l/R)}{p(l/I)}\right)$
[0,0.75)	-0.0328
[0.75,1.5)	0.0062
[1.5,2.25)	0.0678
[2.25,3.0)	0.0729
[3.0,3,75)	0.0031
[3.75,4.5)	-0.0302
[4.5,5.25)	0.0729
[5.25,6.0)	0.0729
[6.0,6.75)	0.0031
[6.75,7.5)	0.0031
[7.5,8.25)	0.0031
[8.25,9.0)	-0.0302
[9.0,9.75)	-0.0497
[9,75,10.5)	0.0729
[10.5,11.25)	-0.0497
[11.25,12.0)	0.0031
[12.0,12.75)	0.0729
[12.75,13.5)	0.0031
[13.5,14.25)	-0.0302
[14.25,15)	0.0031
	$\mathcal{E}(\mathcal{P}, \mathcal{Q}) = 0.2371$

Table 1 From left to right: 1) Set of impact-factor intervals used to define the impact-factor histograms for a given academic institution. 4) Information Gain $\mathcal{E}(\mathcal{P}, \mathcal{Q})$ between impact-factor histograms $\mathcal{P}$ and $\mathcal{Q}$ showed in Fig. 2

example. This same table presents the value of the information gain: $\mathcal{E}(\mathcal{P}, \mathcal{Q}) = 0.2371$.

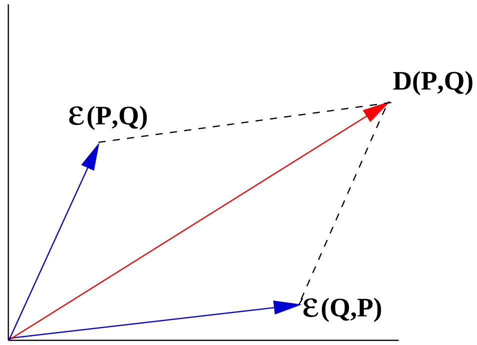

Fig. 3 Computation of the divergence between two academic institutions based on the information gain.

The information gain $\mathcal{E}(\mathcal{P},\mathcal{Q})$ given in Definition 1, is not symmetric with respect to $\mathcal{P}$ and $\mathcal{Q}$, but we may define, following equation (4), a symmetric measure of the divergence between $\mathcal{P}$ and $\mathcal{Q}$ which has this property as well.

Definition 2: Divergence between two academic institutions. The divergence $\mathcal{D}(\mathcal{P},\mathcal{Q})$ is

$$\mathcal{D}(\mathcal{P},\mathcal{Q}) = \mathcal{E}(\mathcal{P},\mathcal{Q}) + \mathcal{E}(\mathcal{Q},\mathcal{P}) \tag{6}$$

with $\mathcal{E}$ as defined in equation (4).

The divergence $\mathcal{D}(\mathcal{P},\mathcal{Q})$ is a symmetric measure of the difficulty of discriminating between the reference academic institution and the input institution: $\mathcal{D}(\mathcal{P},\mathcal{Q}) = \mathcal{D}(\mathcal{Q},\mathcal{P})$. The divergence $\mathcal{D}(\mathcal{P},\mathcal{Q})$ has also the properties of additivity and nonnegativity. Fig. 3 illustrates the computation of the divergence between two academic institutions based on the information gain.

3 Information Conservation Constraint on Specific Fields of Study

After developing the basic axiomatic characterization of a measure of information gain between an input academic institution I and another of reference R, where information gain measures the degree of dissimilarity between these two academic institutions for the total production of universities, we go a step further. In this section we study an approach in which the dissimilarity between two institutions is measured for different scientific fields in the belief that in order to have a clear and more precise picture of institutions' dissimilarities, it is necessary to deepen on specific fields so that we can understand better their relations. This point is stated in the following postulate.

Axiom 4: The effectiveness of the information. *In order to have a clear and more precise picture of institutions' dissimilarities, to deepen on specific fields of study is of primary importance in the comparison of an input academic institution with the reference one so that we can understand better their relations.*

To this aim, the dissimilarity between two institutions can be measured on different scientific fields at which researchers might perceive some significant information following the approach stated in Postulate 5.

Axiom 5. The information conservation constraint. If local information of the reference academic institution in different fields of study is some constraint on the input institution, then *a selective measure of information gain between them involves three steps: (i) specifying interest fields of study in the reference institution ; (ii) for each interest field, zooming in; and (iii) local comparison of the input institution with the reference one.*

This postulate presents the information conservation constraint: properties of the reference academic institution in interest fields of study are equal to the properties of the input institution in the same fields of study. Of course, the problem to be solved is to reformulate the information gain given in Section 2 to process academic institutions in different fields of study.

3.1 Selective Information Gain

Let $Z_1, Z_2, ..., Z_n$ be the interest fields of study for the reference academic institution R; $\mathcal{P}(Z_i)$ and $\mathcal{Q}(Z_i)$ be the local probability distributions of the reference R and the input I computed on the respective fields of study as above.

Thus, $\mathcal{P}(Z_i)$ and $\mathcal{Q}(Z_i)$ are impact-factor histograms which characterize the local information in R and I, respectively. Suppose that every possible observation from $\mathcal{P}(Z_i)$ is also a possible observation from $\mathcal{Q}(Z_i)$.

First, we introduce the local information gain in a particular field of study. To this aim, the information gain in Definition 1 is only applied on publications of this field.

Definition 3: Local information gain. The local information gain between two academic institutions in a field of study Z_i is the information gain between the respective local impact-factor histograms $\mathcal{P}(Z_i)$ and $\mathcal{Q}(Z_i)$ computed using only publications of this field,

$$\mathcal{E}^{Z_i}(\mathcal{P}, \mathcal{Q}) = \mathcal{E}(\mathcal{P}(Z_i), \mathcal{Q}(Z_i)) = \mathcal{U}_{\mathcal{P}}(\mathcal{Q}(Z_i)) - \mathcal{U}_{\mathcal{P}}(\mathcal{P}(Z_i)) \,. \tag{7}$$

with $\mathcal{E}$ as given in Definition 1.

We now introduce a selective measure of information gain between a pair of institutions which are characterized by their respective local probability distributions in different fields of study. This selective information gain sums the local information gain over fields of study.

Definition 4: Selective information gain. The selective information gain $\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P}, \mathcal{Q})$ evaluated on interest fields of study $Z_1, Z_2, ..., Z_n$ of the reference academic institution R is

$$\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) = \sum_{i=1}^{n} \mathcal{E}^{Z_i}(\mathcal{P},\mathcal{Q}) \tag{8}$$

with $\mathcal{E}^{Z_i}(\mathcal{P},\mathcal{Q})$ being the local information gain as given in Definition 3.

The properties of the selective measure of information that we have defined in equation (8) are described as follows.

3.2 Properties of the selective information gain

The first property states the nonnegativity of the selective information measure.

Property 1: Nonnegativity. $\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) \geq 0$, with equality if $\mathcal{P}(Z_i) = \mathcal{Q}(Z_i)$, for $i = 1, 2, \cdots, n$.

Proof.

$$\begin{aligned}
\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) &= \sum_{i=1}^{n} \mathcal{E}^{Z_i}(\mathcal{P},\mathcal{Q}) \\
\text{(by Definition 3)} &= \sum_{i=1}^{n} \mathcal{E}(\mathcal{P}(Z_i),\mathcal{Q}(Z_i)) = \sum_{i=1}^{n} [\mathcal{U}_{\mathcal{P}}(\mathcal{Q}(Z_i)) - \mathcal{U}_{\mathcal{P}}(\mathcal{P}(Z_i))] \\
\text{(by Axiom 3)} &\geq 0, \text{ with equality if } \mathcal{P}(Z_i) = \mathcal{Q}(Z_i), \text{ for } i = 1, 2, \cdots, n
\end{aligned}$$

This proves the nonnegativity property of the selective information gain. This tells us that the overall discrimination information as given by the selective information gain is positive; and there is no discrimination information if the local probability distributions computed on particular fields of study are the same for the reference institution and the input one.

The second property states the additivity of the selective measure of information as follows.

Property 2: Additivity.

$$\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) + \mathcal{E}^{Z_{n+1},\cdots,Z_m}(\mathcal{P},\mathcal{Q}) = \mathcal{E}^{Z_1,\cdots,Z_n,Z_{n+1},\cdots,Z_m}(\mathcal{P},\mathcal{Q})\,.$$

Proof. Following equation (8), this result is trivial.

Additivity of information for independent events is postulated as a requisite property in most axiomatic developments of information theory (Kullback, 1978; Fisher, 1959; Shannon, 1948; Wiener, 1950), and here, Definition 4 states such a fundamental requirement in terms of how local discrimination information determines the overall distinctness between a pair of academic institutions: the overall information obtained from the local gain is the sum of informations yielded by the discrimination using specific fields of study.

The third property relates the measure of information $\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q})$ to the Kullback-Leibler joint information gain of n random variables $Z_1,\cdots,Z_n$.

Property 3: Theorem 2. The selective information gain between $\mathcal{P}$ and $\mathcal{Q}$ of n independent random variables $Z_1,\cdots,Z_n$ defined as given in equation (8), i.e.,

$$\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) = \sum_{i=1}^{n} \mathcal{E}^{Z_i}(\mathcal{P},\mathcal{Q})$$

is the Kullback-Leibler joint information of $Z_1, Z_2,\cdots,Z_n$ up to a nonnegative multiplicative constant:

$$\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q}) = a\,E_{\mathcal{P}}\left[\log\frac{\mathcal{P}(Z_1,Z_2,\cdots,Z_n)}{\mathcal{Q}(Z_1,Z_2,\cdots,Z_n)}\right] \tag{9}$$

with $a \geq 0$ and $E_{\mathcal{P}}$ being the mathematical expectation.

Proof. See Theorem 2 in (Garcia et al., 2001).

The gain $\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P},\mathcal{Q})$, in equation (8), is a selective measure of the information gain of the reference academic institution R from the input institution I.

It follows that the minimum value of this selective information gain between two institutions leads to the most similar academic institutions.

4 Study Case: Benchmarking Spanish Universities in three different fields

4.1 Data source and processing

In order to analyze Information Gain between universities based on their scientific production, the Thomson-Reuters Web of Science databases (SCI and SSCI) were selected as data source. This decision is based on the great regard this database has for research policy makers, as it is considered to store the most relevant scientific literature in the world. Then, we considered 57 Spanish universities, those with at least 150 articles published in the last five years, as the study sample. We selected three different research fields; Information & Communication Technologies (hereafter ICT), Medicine & Pharmacy (hereafter MED) as well as Economics & Business (hereafter ECO) in order to demonstrate the capabilities of the proposed methodology in different contexts as these three areas show different publication patterns and the data source presents a different coverage for each of them. These fields were constructed by aggregating of the Web of Science's subject categories[1]. The selected study time period ranged from 2007 to 2011.

We performed manually a search query for each university in order to download their research output. For this, we used the 'address' filter and took into account all possible naming variants for each institution. Then, we downloaded all citable document types (that is, articles, reviews, notes and letters) and we assigned them

[1] For more information over the construction of the fields of study, the reader is referred to http://www.ugr.es/∼elrobin/rankingsISI_2011.pdf

Field of Study	University analyzed	Position in the field	Nr documents in the field	Type	Foundation
Information & Communication Technologies	Granada	1	558	Public	1526
Medicine & Pharmacy	Navarra	5	1862	Private	1952
Economics & Business	Pompeu Fabra	1	274	Public	1990

Table 2 Sum of the three case studies: fields analyzed and universities

to each university. We only considered as scientific publications those belonging to journals indexed in the Thomson-Reuters Journal Citation Reports (hereafter JCR). These lists of journals are divided by subject category and contain several bibliometric indicators. The most important one is the Impact Factor, which is used to rank journals. This indicator was chosen in order to estimate the impact-factor histograms over which the Information Gain is calculated. However, any other indicator may be used for constructing the histograms such as citations per document, Eigenfactor Scores or any other as long as it is reasonable depending on the purpose of the analysis. Once data was processed, we selected three universities (Granada, Navarra and Pompeu Fabra) which we considered outstanding for each scientific field according to their research output and impact, (Torres-Salinas et al., 2011b), in order to estimate their Selective Information Gain (Definition 4) in relation to the rest of Spanish universities. In fact, these universities are positioned among the top 5 universities for each selected field (Table 2). Also the selected universities are very different institutions, making them interesting cases to analyze. While the University of Granada is a public historical university of a large size and a multidisciplinary focus, the University of Navarra is a private and highly specialized university. On the other hand, the University Pompeu Fabra is a relatively new university with little research output but of great impact.

Finally, we designed an ad hoc benchmarking heliocentric map for each case, allowing the reader to easily analyze the similarity of the study case university with the rest of their scientific field.

4.2 Results

In Figure 4, 5 and 6 we show the results obtained for our case study. In these figures, the reference university is positioned in the middle of the heliocentric map (Granada, Navarra and Pompeu Fabra) and the other 30 universities are placed around it considering their similarity. These universities are the ones with the lowest values of selective information gain (that is, the highest values of similarity). Universities are ordered clockwise considering their positions in their national rankings (http://rankinguniversidades.es) starting on the top of the figure. This way, similarity can be perceived also considering the 'best' or 'worst' universities.

In Figure 4 we observe that the most similar universities to the University of Granada in the field of ICT are the polytechnic universities (that is, Cataluña, Valencia and Madrid) along with Carlos III, Málaga and País Vasco. But there are also other universities in top positions in their national rankings which are less alike such as Pompeu Fabra, Barcelona or Santiago. The reason for this is that these three universities have 49%, 45% and 40% respectively of their total output in journals positioned in the first quartile of their subject category, while Granada and the other universities alike have only around 30% (Cataluña) and 34% (Granada). Therefore, their publication profiles are well portrayed considering their similarity (Selective Information Gain). On the other side, we find that those

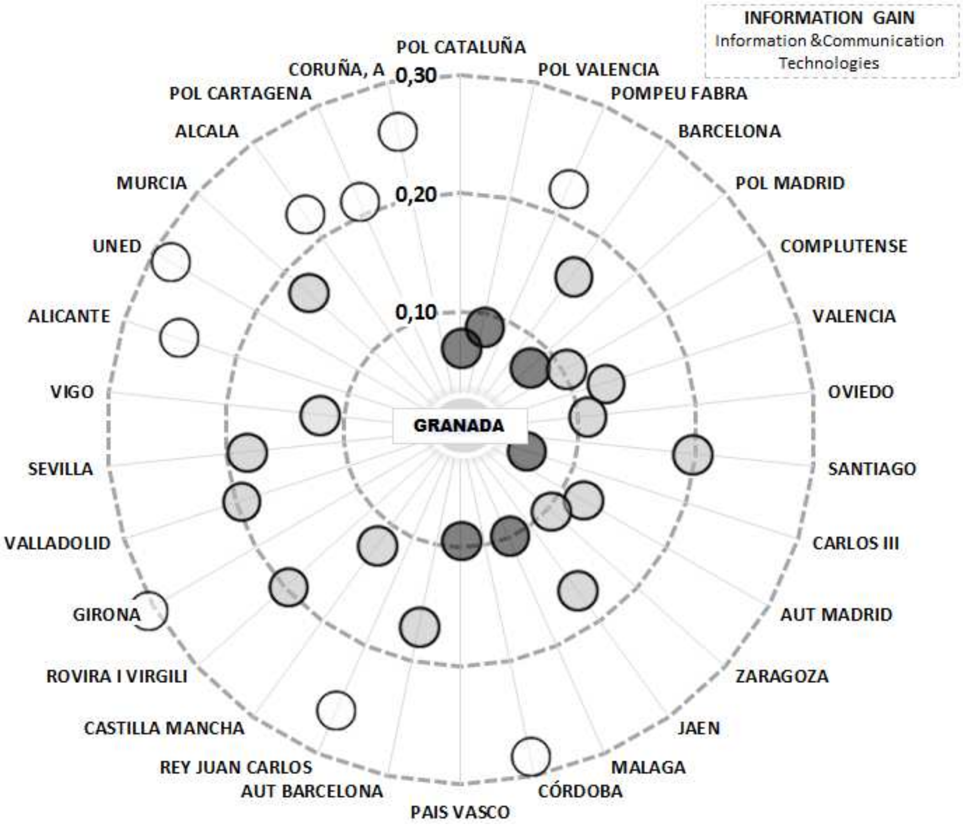

Fig. 4 Heliocentric map representing the Selective Information Gain for University of Granada in the field of 'Information & Communication Technologies'. Period 2007-2011.

ranked in the lowest positions of the national ranking are the ones with higher information gain, and consequently, more dissimilar.

Figure 5 represents the case study of Navarra for the field of MED. In this case, we observe that the configuration has the form of a spiral, where the most alike universities are also those situated in the top of the national ranking and as we go down in the ranking, universities are more dissimilar. The most similar universities to Navarra, i.e., Barcelona, Autonoma de Barcelona, Autonoma de Madrid, Valencia and Salamanca, have several characteristics in common; they have all produced more than 1000 documents for the study time period, they have similar percentages of output published in journals belonging to the first quartile of their category and they are all above the national average which is between 41% for Valencia and 52% for Barcelona. They therefore perform very similarly not just considering their total output but also their publication profile, as it is shown by their Information Gain indicator. Probably, this can be explained by analyzing the nature of the university of Navarra, which is private, and follows a more specific or rigorous research policy in this scientific field than the public ones, therefore focusing its publication profile in excellence and copying successful models.

Finally, in Figure 6 we show the case of Pompeu Fabra in the field of ECO. In this case, due to the high impact of its publications, there are not many universities similar to its profile. This anomalous pattern compared to the rest of the Spanish universities has already been suggested (Garcia et al., 2012a). Considering their Selective Information Gain, only two universities have values equal or lower to 0.18, these are Autonoma de Barcelona and Carlos III; both showing similar percentages of publications in the first quartile, 32% and 35% respectively. Interestingly, Autonoma de Barcelona, Pompeu Fabra, Carlos III and Alicante coincide as top Spanish universities in this field by previous studies (Dolado et al., 2003; Lubrano et al., 2003).

However, the ones with the highest percentages of publications in the first quartile are Navarra and Pompeu. Although one would think of their similarity, according to Figure 6 this does not occur. The main reason for this has to do with the great level of specialization of both institutions in different specialities. While Pompeu Fabra is highly specialized in Economics, Navarra is specialized in Management, mainly due to its link with the IESE Business School. We must take

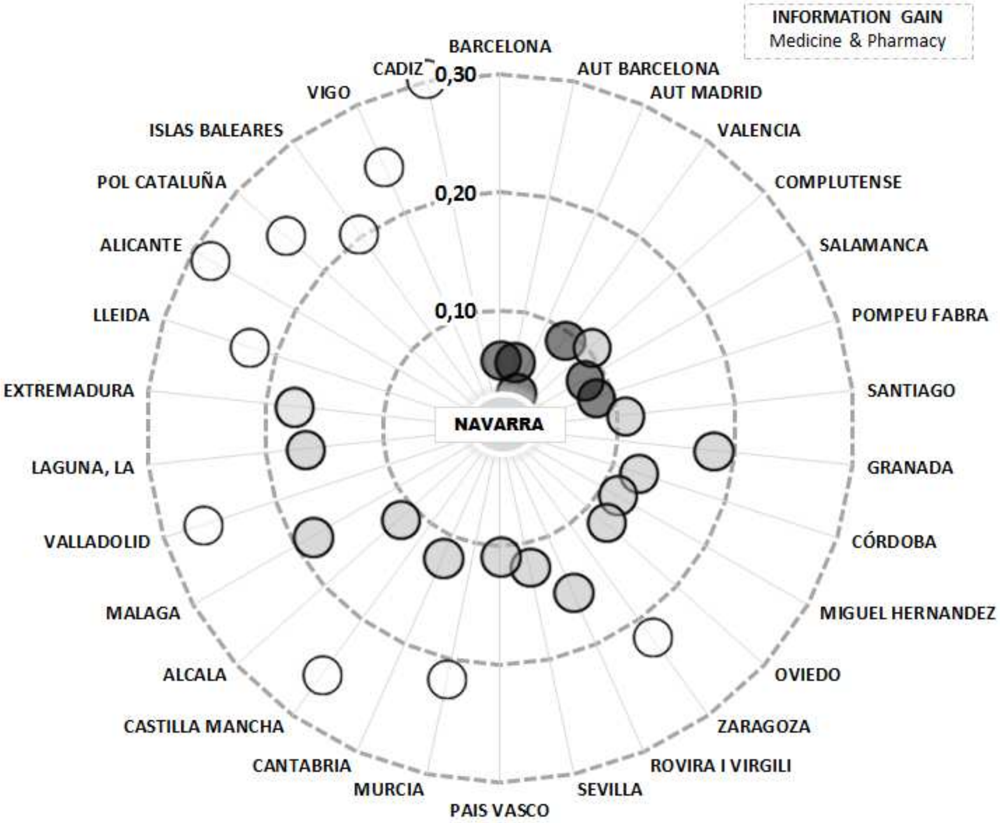

Fig. 5 Heliocentric map representing the Information Gain for University of Navarra in the field of 'Medicine & Pharmacy'. Period 2007-2011.

into account that the Selective Information Gain indicator uses the Impact Factor as a reference and that the quartile thresholds are different for each category. While in Economics there are 75 journals ranked in the first quartile in a range between 7.432 and 1.297, in Management there are just 36 journals, ranging their Impact Factor in first quartile journals from 6.720 to 2.473.

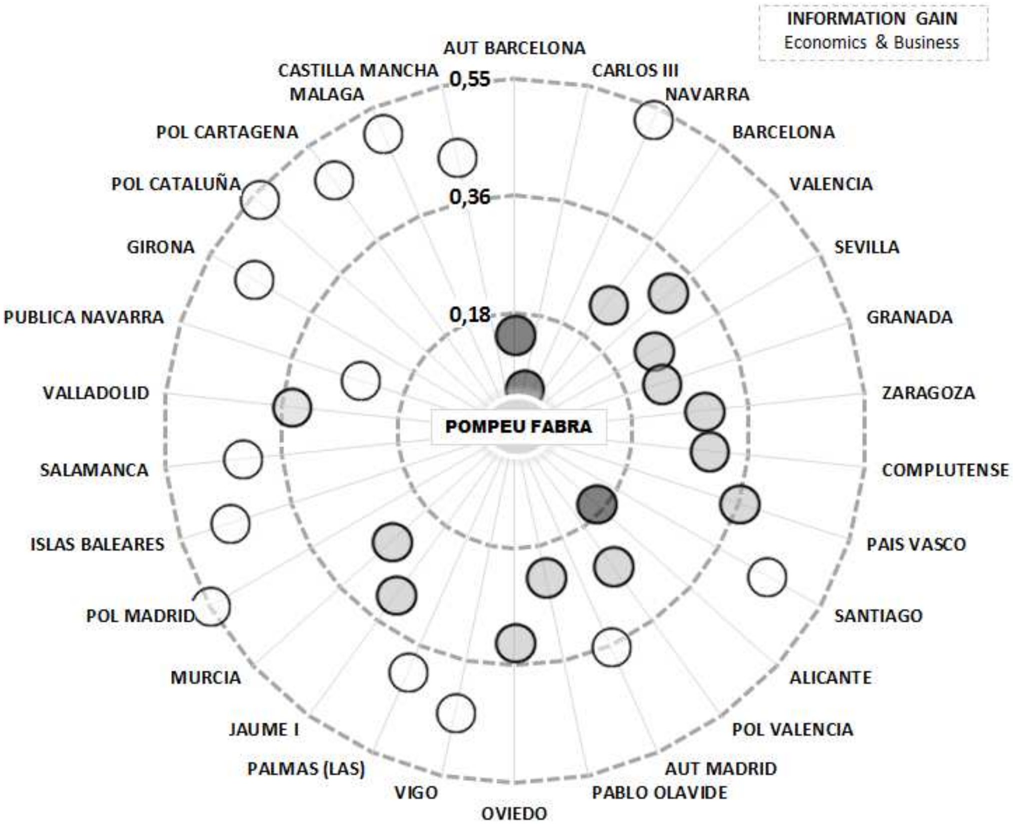

Fig. 6 Heliocentric map representing the Information Gain for University of Pompeu Fabra in the field of 'Economics & Business'. Period 2007-2011.

4.3 Validation of the Selective Information Gain

In order to validate the Selective Information Gain indicator, we compare it with other indicators also aimed at measuring research output similarities between two subjects. We use the obtained Selective Information Gain indicators based on impact-factor histograms that characterize probabilistically the research output structure of the universities. In this case, we will compare them in the three scenarios above mentioned with the University-University Similarity matrix based on the Journal Publication Profile proposed by (Garcia et al., 2012a). This method-

University-university similarity based on their journal publication profile
1. Obtain list of journals on which each institution has published for the study time period
2. Apply weights to journals for each institution.
3. Construct a journal-by-institution matrix.
4. Extract values from an institution-institution matrix.
5. Apply a second-order approach to emphasize similarities among institutions.
6. Perform a complete linkage clustering method in order to set the institutions groups according to their journal publication profile.

Table 3 Sum of the methodology for mapping universities according to their journal publication profile

ology is based on the idea that two universities are similar when they both publish in the same journals. In Table 4 we present the methodological procedure for obtaining such profile.

However, the Selective Information Gain is a measure of dissimilarity which makes it unfeasible to compare it directly with the university-university similarity matrix, therefore we take the Selective Information Gain values from the lowest to the highest one in other to make them comparable. We have used the Pearson and Spearman correlation coefficients for validating our measure. Results are shown in Table 4. Both measures offer similar results when studying the similarity between research outputs of academic institutions. Clearly, results are more consistent in MED, followed by ICT and lastly, ECO, where it reaches the lowest values (being in any case greater than 0.7).

5 Conclusion

In this paper we present four theoretic information measures for benchmarking research performance applying it in a case study in academic institutions based

University	**Granada**	**Granada**	**Navarra**	**Navarra**	**Pompeu Fabra**	**Pompeu Fabra**
Field	**ICT**	**ICT**	**MED**	**MED**	**ECO**	**ECO**
Coefficient	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
Correlation	0.7602	0.8414	0.9339	0.8399	0.7135	0.7490

Table 4 Correlation coefficients between the Selective Information Gain and the university-university similarity based on journal publication profiles

on the dissimilarity between the research outputs of universities. We illustrate their usefulness by applying them to the impact-factor histograms in three case studies in which we compared a given Spanish university with the rest of them in a given research field. The chosen universities were Granada, Navarra and Pompeu Fabra for ICT, MED and ECO respectively as they are universities with different institutional profiles which outstand for each of the chosen scientific fields. From the four measures described in Section 2 and 3 we applied the Selective Information Gain measure (Definition 4) as we consider it the most suitable one.

From the experimental results presented in this paper, the following five postulates were found to be relevant for benchmarking research performance by means of the difference between impact-factor histograms for the reference and input academic institutions:

- **Postulate 1.** A measure of how unexpected the single event ”a publication with impact factor in certain interval occurs” was, depends only upon its probability.
- **Postulate 2.** An estimate of how unexpected the research performance of a reference academic institution was from some probability distribution is simply defined as the mathematical expectation of how unexpected its single events were from this distribution.

- **Postulate 3.** The research-output of a reference academic institution is more unexpected from an "estimated" distribution than from the "true" distribution.
- **Postulate 4. The effectiveness of the information.** In order to have a clear and more precise picture of institutions dissimilarities, to deepen on specific fields of study is of primary importance in the comparison of an input academic institution with the reference one so that we can understand better their relations.
- **Postulate 5. The information conservation constraint.** If local information of the reference academic institution in different fields of study is some constraint on the input institution, then a selective measure of information gain between them involves three steps: (i) specifying interest fields of study in the reference institution; (ii) for each interest field, zooming in; and (iii) local comparison of the input institution with the reference one.

The selective gain is a measure of information gain between two academic institutions, such that satisfies Postulates 1–5. Also it has the property of additivity.

The main result of this study is that the selective information gain measure relates closely to similarity between universities as perceived by using different models. In order to validate such information measure we have applied two correlation coefficients which show the high coherence between the results obtained in the present study with those from previous ones (Garcia et al., 2012a,b).

In conclusion, both theoretical and empirical results imply that it can be used to benchmark research performance of academic institutions using an information theoretic approach.

Although in this study we have applied the Information Gain to universities, it could also be used to compare the research output and impact of other research units (e.g., researchers, countries) and could also be combined with other bibliometric indicators such as the number of citations.

Acknowledgments. This research was sponsored by the Spanish Board for Science and Technology (MICINN) under grant TIN2010-15157 cofinanced with European FEDER funds. Nicolás Robinson-García is currently supported by a FPU grant from the Spanish Ministerio de Educación y Ciencia. Thanks are due to the reviewers for their constructive suggestions.

References

Adams, J., Gurney, K. & Marshall, S. (2007). Profiling citation impact: a new methodology. Scientometrics, 72(2), 325-344.

Abramo, G., DAngelo, D.A. & Di Costa, F. (2011). National research assessment exercises: a comparison of peer review and bibliometric rankings. Scientometrics, 89(3), 929-941.

Aczél, J., and Daróczy, Z. (1975). On measures of information and their characterizations. Volume 115 in Mathematics in Science and Engineering. Academic Press, New York.

Burbea, J., and Rao, C.R. (1983). On the convexity of divergence measures based on entropy function. IEEE Trans. Inf. Theory, 28(3), 489-495.

Carpenter, M. P., Gibb, F., Harris, M., Irvine, J., Martin, B. R. & Narin, F. (1988). Bibliometric profiles for British academic institutions: an experiment to develop research output indicators. Scientometrics, 14(3-4), 213-233.

Dolado, J. J., García-Romero, A. & Zamarro, G. (2003). Publishing performance in economics: Spanish rankings (1990-1999). Spanish Economic Review, 5, 85-100.

Fisher, R.A., (1950). The logic of inductive inference. J. Roy. Statist. Soc., Vol. 98, 39-54; Contributions to Mathematical Statistics, John Wiley and Sons, New York, paper 26.

Garcia, J.A., Fdez-Valdivia, J, Fdez-Vidal, Xose R., and Rodriguez-Sanchez R. (2001). Information Theoretic Measure for Visual Target Distinctness. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(4), 362-383.

García, J. A., Rodríguez-Sánchez, R., Fdez-Valdivia, J., Robinson-García, N. & Torres-Salinas. (2012a). Mapping academic institutions according to their journal publication profile: Spanish universities as a case study. Journal of the American Society for Information Science and Technology. Accepted for publication.

García, J. A., Rodríguez-Sánchez, R., Fdez-Valdivia, J., Torres-Salinas, D. & Herrera, F. (2012b). Ranking of research output of universities on the basis of the multidimensional prestige of influential fields: Spanish universities as a case of study. Scientometrics, DOI: 10.1007/s11192-012-0740-7.

Havrda, J.H., and Charvat, F. (1967). Quantification method of classification processes: Concept of structural α-entropy'. Kybernetika, Vol. 3, 30-35.

Hazelkorn, E. (2011). Rankings and the reshaping of higher education: the battle for world-class excellence. Basingstoke, New York: Palgrave-MacMillan.

Jeffreys, H. (1946). An invariant form for the prior probability in estimating problems. Proc. Roy. Soc. London, 186A, 453-461.

Kapur, J.N. (1984). A comparative assessment of various measures of directed divergence. Advances Manag. Stud., Vol. 3, 1-16.

Kullback, S. (1978). Information theory and statistics. Gloucester, Mass. Peter Smith.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency.' Ann. Math. Statist., 22, 79-86.

Lubrano, M., Bauwens, L., Kirman, A. & Protopopescu, C. (2003). Ranking economics departments in Europe: A statistical approach. Journal of the European Economic Association, 1, 1367-1401.

Lundberg, J. (2006). Bibliometrics as a research assessment tool impact beyond the impact factor [PhD dissertation]. Stockholm: Karolinska Institutet.

Moed, H. F. (2008). UK research assessment exercises: Informed judgements on research quality or quantity? Scientometrics, 74(1), 153-161.

Ortega, J. L., Lopez-Romero, E. & Fernandez, I. (2011). Multivariate approach to classify research institutes according to their outputs: The case of the CSIC's institutes. Journal of Informetrics, 5(3), 323-332.

Rao, C.R. (1982). Diversity and dissimilarity coefficients: A unified approach. Theoretic Population Biology, Vol. 21, No. 1, 24-43.

Renyi, A. (1961). On measures of entropy and information. Proc. Fourth Berkeley Sump. Math. Stat. and Prob., University of California Press, Vol. 1, 547-561.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Tech. J., Vol. 27, 379-423; 623-656.

Sharma, B.D., and Mittal, D.P. (1977). New non-additive measures of relative information. Journ. Comb. Inf. Syst. Sci., Vol. 2, 122-132.

Shin, J. C. (2009). Classifying higher education institutions in Korea: a performance-based approach. Higher Education, 57(2), 247-266.

Tijssen, R. J. W., van Leeuwen, T. N. & van Wijk, E. (2009). Benchmarking university-industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities. Research Evaluation, 18(1), 13-24.

Torres-Salinas, D., Moreno-Torres, J. G., Delgado-Lopez-Cozar, E. & Herrera, F. (2011a). A methodology for Institution-Field ranking based on a bidimensional analysis: the IFQ2A index. Scientometrics, 88(3), 771-786.

Torres-Salinas, D., Moreno-Torres, J. G., Robinson-García, N., Delgado-López-Cózar, E., Herrera, F. (2011b). Rankings ISI de las Universidades Españolas según campos y disciplinas científicas (Second ed. 2011). El Profesional de la Información, 20(6), 701-709.

Turner, D. (2007). Benchmarking universities: league tables revisited. Oxford Review of Education, 31(3), 353-371.

Vanclay, J. K. & Bornmann, L. (2012). Metrics to evaluate research performance in academic institutions: a critique of ERA 2010 as applied in forestry and the indirect H2 index as a possible alternative. Scientometrics, 91(3), 751-771.

Wiener, N. (1950). The Human use of human beings, Houghton Mifflin Co., Boston.

A Appendix: Forms of Gains and Divergences

The simplest form of gain $\mathcal{E}(\mathcal{P},\mathcal{Q})$ and divergence $\mathcal{D}(\mathcal{P},\mathcal{Q})$ is based on Axioms 1, 2, and 3 (Section 2). By Theorem 1, the form of the information gain $\mathcal{E}(\mathcal{P},\mathcal{Q})$ is (up to a nonnegative multiplicative constant):

$$\mathcal{E}(\mathcal{P},\mathcal{Q}) = \sum_{l} p(l/R) \log \frac{p(l/R)}{p(l/I)} \tag{10}$$

with $\mathcal{P} = \{p(l/R)\}$, and, $\mathcal{Q} = \{p(l/I)\}$; $p(l/R)$ and $p(l/I)$ being the probability of occurrence of a publication with impact factor in the range $[l, l + \Delta l)$, for the reference institution R and the input institution I respectively.

Substituting equation (10) in equation (6), the form of the divergence $\mathcal{D}(\mathcal{P}, \mathcal{Q})$ is (up to a nonnegative multiplicative constant):

$$\mathcal{D}(\mathcal{P}, \mathcal{Q}) = \sum_{l} [p(l/R) - p(l/I)] \log \frac{p(l/R)}{p(l/I)}. \tag{11}$$

By adding a postulate relative to the effectiveness of the information as well as the information conservation constraint to the axiomatic characterization (Section 3), we obtain the selective gain $\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P}, \mathcal{Q})$. By Theorem 2, the selective gain between R and I is similar to the Kullback-Leibler joint information gain of n interest fields of study $Z_1, Z_2, \cdots, Z_n$ between the reference institution and the input one I. The selective gain is (up to a nonnegative multiplicative constant):

$$\mathcal{E}^{Z_1,\cdots,Z_n}(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{n} \sum_{l} p(l/R_{Z_i}) \log \frac{p(l/R_{Z_i})}{p(l/I_{Z_i})} \tag{12}$$

with $Z_1, \cdots, Z_n$ being the interest fields of study of the reference academic institution R; $(p(l/R_{Z_i}))_l$ being the local probability distribution computed on field of study Z_i for the reference institution R; $(p(l/I_{Z_i}))_l$ being the local distribution computed on Z_i for the input institution I.