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

Self-defined information indices: application to the case of university rankings

A. Ferrer-Sapena[1] · **E. Erdogan** · **E. Jiménez-Fernández** · **E.A. Sánchez-Pérez** · **F. Peset**

Received: date / Accepted: date

Abstract University rankings are now relevant decision-making tools for both institutional and private purposes in the management of higher education and research. However, they are often computed only for a small set of institutions using some sophisticated parameters. In this paper we present a new and simple algorithm to calculate an approximation of these indices using some standard bibliometric variables, such as the number of citations from the scientific output of universities and the number of articles per quartile. To show our technique, some results for the ARWU index are presented. From a technical point of view, our technique, which follows a standard machine learning scheme, is based on the interpolation of two classical extrapolation formulas for Lipschitz functions defined in metric spaces—the so-called McShane and Whitney formulae—. In the model, the elements of the metric space are the universities, the distances are measured using some data that can be extracted from the Incites database, and the Lipschitz function is the ARWU index.

A. Ferrer-Sapena ([1] Corresponding Author)
IUMPA, Universitat Politècnica de València, 46022 Valencia, Spain
Tel.: +34-963877663
Fax: +34-963877669
E-mail: anfersa@upv.es

E. Erdogan
Marmara University, Istanbul, Turkey
E-mail: ezgi.erdogan@marmara.edu.tr

E. Jiménez-Fernández
Universitat Jaume I, Castellón, Spain
E-mail: jimeneze@uji.es

E.A. Sánchez-Pérez
IUMPA, Universitat Politècnica de València, 46022 Valencia, Spain
E-mail: easancpe@mat.upv.es

F. Peset
IUMPA, Universitat Politècnica de València, 46022 Valencia, Spain
E-mail: mpesetm@upv.es

Keywords Reinforcement learning · Metric space · Lipschitz extension · Shanghai · ARWU · University ranking
JEL C60 C61 C88

Mathematics Subject Classification (2000) 54E35 · 26A16

1 Introduction and basic definitions

University rankings are usually developed using some specific indexes that consider relevant information in relation to different aspects of academic activity. The social influence of these rankings has been deeply analysed in the last few years, and also the main technical aspects of the comparison between them (Aguillo et al. 2010; Chen and Liao 2012) which has become an interesting source of analysis on the influence of the main cultural areas and scientific regions of the world (Saisana et al. 2011). Furthermore, it is clear today that they play a central role in the design of policies affecting issues as different as national scientific research programmes, library policies, university funding and education policies, and many others (Lim and Øerberg 2017; Margison 2014; Pagell 2014). However, indices that provide rankings are often calculated for a selected group of institutions, for which all the values of the specific variables are known. This makes the definition of the indices and their use a "vicious cycle": The "best" universities are taken to choose the variables that determine the definition of the indexes, which show that these universities are, in fact, the best. This makes it reasonable to ask for procedures to increase the set of institutions for which the indices can be computed, allowing for comparisons and redefinitions in some cases. We present in this paper a method for assessing larger sets of institutions for which the indices would also make sense, under the restriction of not knowing the value of the original set of variables selected for a (potentially large) part of the new set.

Among others, the Academic Ranking of World Universities (known as ARWU ranking or Shanghai ranking) is an important reference for the worldwide comparison of institutions involved in higher education and scientific research. However, the score in which it is based is not computed for a large set of institutions. In particular, in the Incites database it can be obtained bibliometric information for a lot of universities not appearing in the ARWU ranking, for which the related score is not known. It must be taken into account that for the computation of combined indices there are a lot of different sources of information that are considered, that sometimes simply do not make sense for institutions of a different class. However, similarity among universities —measured using for example a metric based in the number published papers in different quartiles of the JCR and total number of citations—, can allow to provide an expected value of the ARWU score for universities that are out of this ranking. In order to get this, we need a metric space (a set D of universities together with a metric d based in this kind of similarity relations), an index that is known for a meaningful subset of D (in our case, the ARWU score for a subset of top universities), and an extrapolation method.

Thus, the purpose of this paper is twofold. First, we are interested in presenting a new method for extending specific indices to larger classes of entities, which is ob-

tained by applying some classical results on the extrapolation of real Lipschitz functions to perform a new machine learning procedure. The result is a typical reinforcement learning algorithm based on classical extension theorems for real functions—the McShane-Whitney Theorem—in a new mathematical environment. Second, we apply this technique to provide a new tool to address the problem that actually motivated the mathematical part of our research: the use of prestige-based indices to build university rankings that include institutions of different sizes and characteristics. Although there are many studies on university classifications, there are not many published tools that cover the objective of the algorithm provided here, which is the extrapolation of university rankings from known to unknown situations. However, some other authors have already analyzed this from this point of view, see for example the remarkable contribution provided in Tabassum et al. 2017.

Potential applications of our algorithm are easy to find. The main one, as we said above, is essentially to extend the definition of indexes to larger sets. For example, it can be used for comparison among different indexes that are computed for different sets of institutions, which could improve the results of the productive comparative analysis of rankings (Aguillo et al. 2010; Chen and Liao 2012; Cinzia and Bonaccorsi 2017; Kehm 2014). It could also help to correct the negative effects on the rankings of the native language of the countries in which the universities are located and wrong citation counts motivated by non-anglosaxon names (Van Raan et al. 2011), by extrapolating the associated scores using metrics which do not use these variables. Finally, it can be used for getting specific estimates of prestigious scores for almost all universities of the world –whenever some bibliometric date are known—, which can help all the institutions to measure the comparison with bigger and more powerful universities of other countries.

We will use Incites as main source of bibliometric indicators for a large set of universities. We will center the definition of our metric in a metric completely based in number of published papers and citations. The influence of these variables in the university rankings has been analyzed since university rankings appeared, and is nowadays well-known (see Luo et al. 2018 and the references therein; see also Cancino et al. 2017). Some contributions have also been made on how rankings are defined (multi-attribute rankings) and how universities can develop optimal strategies for scaling them using only the mathematical properties of the underlying index structure (Bougnol and Dulá 2013). Rankings are based on indices, and the indices are supported by models with particular mathematical structures. (More examples of rankings based on multivariable indices can be found in U-Multirank 2019.)

Let us briefly present our method. Consider a set of D from entities such as journals, authors, libraries, or universities for which you want to have an index-based evaluation model. Assume that there is a metric d that measures the similarity of every two elements in D. Suppose that a "quality" index I is defined over a subset $D_0 \subseteq D$, and is coherent with the metric d, that is, if the distance among two elements a and b is small, then the values of $I(a)$ and $I(b)$ are similar. Using a machine learning scheme based on extrapolation techniques for Lipschitz functions, we can extend the index I to the whole set D. This gives a class of explicit formulas for calculating the index I for entities in the complementary set $D_1 = D \setminus D_0$, where it was not originally defined. In a second step, we use a reinforcement learning method for

choosing the best formula in the class with the aim of computing an approximation to I in D_1. Since the formula for computing such an extension depends on I—that is defined only in D_0—and on the metric d, we call to such extended function a self-defined quality index.

We designed this method to face the problem of how to define in a fair and correct way a ranking of universities in a group that contains for example small institutions that cannot be measured with the same standards as the big ones. The idea is to try to avoid that strong requirements—such as, for example, having Nobel Prizes—immediately exclude some good (but small) centers from having good positions in the ranking, by creating a similar but more inclusive method for computing it.

Our ideas will be presented in four sections. After the some preliminaries, we will explain in Section **??** the mathematics concerning the procedure and the algorithm itself. Section **??** will show the concrete application for defining a self-defined index. Finally, in Section **??** we will show how to apply the model to the problem explained above.

Let us introduce now some basic definitions that will be used throughout the paper. Let $\mathbb{R}^+$ be the set of non-negative real numbers. A *distance* in a set D is a function $d : D \times D \to \mathbb{R}^+$ such that for $a,b \in D$,

(i) *Separation:* $d(a,b) = 0$ if and only if $a = b$,
(ii) *Symmetry:* $d(a,b) = d(b,a)$, and
(iii) *Triangle inequality:* if $c \in D$, then $d(a,b) \leq d(a,c) + d(c,b)$.

Such a function is also called a metric on D. Although we restrict our attention in the present paper to weighted Euclidean distances, there are a lot of different metrics that can be used to model the problem that we face here (see for example Deza and Deza 2009.)

A real valued function acting in a metric space (D,d) is said to be Lipschitz if it satisfies the inequality $\left|f(a) - f(b)\right| \leq K\,d(a,b)$ for a certain constant $K > 0$ and for all $a,b \in D$. The Lipschitz constant of f is the infimum of all the constants K above. Often we will use the same symbol K for this optimal constant. The *McShane-Whitney Theorem* states that for every subspace B of a metric space (D,d), and every Lipschitz function $f : B \to \mathbb{R}$ with Lipschitz constant K, there exists an extension $\hat{f}$ of f to D such that $\hat{f}$ is also a Lipschitz function with the same Lipschitz constant K (see for example Th.4.1.1 in Cobzaş et al. 2019; see also Section 5.2 in this book).

Two of all possible extensions are in a certain sense canonical, and are given by the following formulas,

$$f^M(b) := \sup_{a\in B}\{f(a) - K\,d(b,a)\}, \quad \text{and} \quad f^W(x) := \inf_{a\in B}\{f(a) + K\,d(b,a)\},$$

that are defined for all $b \in D$ and equal to f if $b \in B$. They are called the McShane extension and the Whitney extension, respectively. It is easy to see that convex combinations of these formulas are also extensions of f to D. We will use such type of extensions in the present paper.

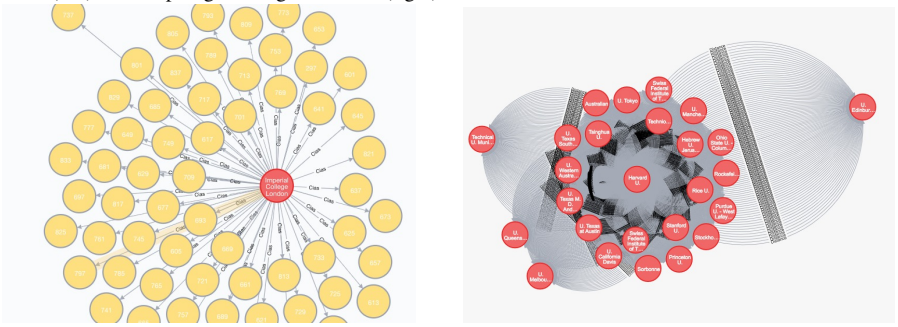
2 The model: distances versus indices

Based on some recent developments in reinforcement learning, we present in this chapter a new mathematical framework for producing automatically corrected general indices from their definition in a given subset in which the index is clear and correctly defined. An iterative procedure is proposed, taking into account the original subset, as well as the corrections obtained by contrasting with other data from new sets of information. The general framework, that has been originally developed for financial time series, has been presented recently in the paper Falciani et al. 2020, to which the interested reader is referred; see also the references therein for an update of the required mathematical tools for this reinforcement learning method of artificial intelligence. More information on the use of Lipschitz functions for reinforcement learning algorithms—often based on metric graphs—, can be found in Asadi et al. 2018; von Luxburg and Bousquet 2004; Rao 2015. A similar research purpose, although based on different mathematical tools, can be found in Tabasumm et al. 2017; see also Çakır et al. 2015; Dobrota et al. 2016; Rosa et al. 2012, for other analytical procedures.

In general, it seems difficult to define the properties of a set of entities in Information Science —journals, scientists, institutions, editorials,...— that make it possible to characterize the role in the entities in an analytical model. This is the first step for a rigurous analysis, and must be carefully done. For example, suppose that we are interested in analysing an impact-based system for research production evaluation of research institutions that uses the distribution by quartiles of the index SNIP from Scopus. For such a model, a 4-coordinates vector is enough for indentifying a given institute, writing in each coordinate the number of papers published in journals in each quartile of the list of a previously fixed year.

However, to define the relevant variables to characterize an entity in the model is not the aim of the present paper, in which we assume that the Information Scientist has already developed a method for determining variables that must be taken into account. In any case, our modelling of the problem starts by the definition of the metric space to which all the entities that are considered as elements of our analysis belong to. In Figure **??** the reader can find a representation of the distances from a given university (Imperial College, London) to the whole set of universities (left), and also of the topological neighbourhood of the Harvard university in the model (right), made using the graph-database platform Neo4j.

Fig. 1 Graph-based representation of distances and neigbourhoods in the model: distances among universities (left) and a topological neighbourhood (right).



The main object of our model is a metric space (D,d) together with an index I, which will define a triplet (D,d,I) that we call a metric-index model. Sometimes, the index I is only defined for a subset of elements of D —say, D_0—; in this case, an extension of such a function preserving the relation between d and I can be obtained, and we say in this case that the extended I is a self-defined index. This is exactly the situation that we are interested in studying in the present paper. The procedure to extend the values of I to the whole metric space have to preserve some basic properties of I. The main one is the Lipschitz constant, that represents the relation among I and the metric d, that is, how far a strong proximity relation among elements $a,b \in D$—that is, a small value of $d(a,b)$—, implies that the corresponding values $I(a)$ and $I(b)$ have to be similar too—that is, the difference $|I(a)-I(b)|$ has to be small—. To control this is the reason why we introduce below the notion of coherence.

Definition 1 Let $K > 0$. An index $I : (D,d) \to \mathbb{R}^+$ is K-coherent if it satisfies the Lipschitz inequality for the constant K. That is,

$$\big|I(a)-I(b)\big| \leq K\, d(a,b) \quad \text{for all} \quad a,b \in D.$$

We will say that K is the coherence constant of I if the infimum of all constants K' satisfying this property is equal to K, that is, K is the Lipschitz constant of I.

In the case of the analysis of the ARWU university ranking that we will present further in the paper, we will use this notion to measure how appropriate a metric is to model a given previously defined index, that will be in our case the ARWU score. Both the McShane and the Whitney formulae—that will be used in the extrapolation formula that provides the self-defined index—, preserve the coherence (that is, the Lipschitz constant).

3 Extending an information index from a field D_0 to a field D_1

As we have explained, a metric-index model (D,d,I) is good if the relationship between the values of I and the distance d that describes the affinity of the elements of D is also good. Therefore, in the model, two elements a and b of D are "similar"

if $d(a,b)$ is small, and in this case the values of the index I —which is supposed to summarily describe a "rating" of these elements—, have to be similar as well. In this case, I is K-coherent —for a value K that is intended to be small— with respect to the control distance d.

Let D_0 and D_1 be a partition of D. Let us describe the formal way of analyzing the following **problem:** *we want to know if a suitable extension of an index I, that is K-coherent with respect to d for a given field D_0 and with small constant K, can also be considered as K-coherent with the same constant K in D_1 to which I is extended.* Remember that the basic assumption is that $D_0 \cup D_1 = D$ is a metric space with a (common) distance d defined in it. It is asumed that d describes in both D_0 and D_1 the similarity of two elements.

In formal terms, the above problem can be described as follows. Suppose that some expected values of the index I are known only for a subset S_1 of elements of D_1, although I is fully known in D_0. Thus, given an index that is defined in D_0 and is K-coherent with small K, can I be defined in D_1 as a Lipschitz extension of I to D with the same constant K? Several methods can be used to solve this problem. In this paper we propose a new one, based on the similarity with a particular class of Lipschitz extensions, provided in this case by the family of convex combinations of the McShane and the Whitney extensions of I in D_0.

The method follows the next steps.

(1) Fix an index $I : D_0 \to \mathbb{R}$. The Lipschitz constant of I in D_0 is K. Consider the family
$$L_{ext} := \left\{ I_\alpha^{M,W} = \alpha I^M + (1-\alpha) I^W : 0 \le \alpha \le 1 \right\},$$
where I^M and I^W are the McShane and Whitney extensions of I, respectively.

(2) Assume that the index I is known for all the elements of D_0, in which —due to the hypothesis of the method— I is defined, and it is understood to be a good model for the property it reflects.

(3) In principle, the function I is not supposed to be known in D_1, although it is assumed that it could be defined and Lipschitz. However, its Lipschitz constant is not known. Some information on the expected value of I in D_1 is supposed to be known. Concretely, the expected value of I is known for the elements of a sample $S_1 \subseteq D_1$.

(4) Now, we can calculate the best α using a least squares procedure that will give the best extension of I in the set L_{ext}. That is, we compute a value of α_1 for which
$$\min_{0\le\alpha\le 1} \sum_{x\in S_1} \left| I(x) - I_\alpha^{M,W}(x) \right|^2 = \sum_{x\in S_1} \left| I(x) - I_{\alpha_1}^{M,W}(x) \right|^2.$$
This expression has the meaning of an error, an therefore also gives an idea of the extent to which $I_{\alpha_1}^{M,W}$ is a canonical extension of I belonging to the family of convex combinations L_{ext}.

(5) The function $I_{\alpha_1}^{M,W}$ is a **best extension** of I_0 to D_1, preserving the K-coherence of the original index I when extended to D_1.

Note that the best extension computed depends on the sample, so it could change when the size of S_1 increases. The larger the S_1, the better the approximation $I_{\alpha_1}^{M,W}$.

When implementing an iterative process by increasing the size of S_1 at each step, we have a typical machine learning/reinforcement learning scheme for improving the fit of the extension of I to D.

Remark 1 Note that the coherence constant K is preserved in our extension. Indeed, it is well known that the McShane and the Lipschitz extensions preserves the Lipschitz constant K. Then, for any $\alpha \in (0,1)$ we have that

$$\begin{aligned}\left|I_\alpha^{M,W}(a) - I_\alpha^{M,W}(b)\right| &= \left|\alpha I^M(a) + (1-\alpha)I^W(a) - \alpha I^M(b) - (1-\alpha)I^W(b)\right| \\ &\leq \alpha\left|I^M(a) - I^M(b)\right| + (1-\alpha)\left|I^W(a) - I^W(b)\right| \leq K\,d(a,b).\end{aligned}$$

In general, we can choose the convex combination of I^M and I^W depending on the problem, and the analyst can use her/his own experience to give a reasonable value to α. However, in the case we consider we have a supervised algorithm —that is, we complete our algorithm with the minimization of the error of the extension for a given sample set—, and so an explicit formula for α can be obtained, which is given below.

Remark 2 The value $0 \leq \alpha_1 \leq 1$ that attains the minimun

$$\min_{0\leq\alpha\leq 1} \sum_{x\in S_1} \left|I(x) - I_\alpha^{M,W}(x)\right|^2 = \sum_{x\in S_1} \left|I(x) - I_{\alpha_1}^{M,W}(x)\right|^2$$

is given by

$$\alpha_1 = \frac{\sum_{x\in S_1}\left(I_0^W(x) - I_0^M(x)\right)\left(I_0^W(x) - I(x)\right)}{\sum_{x\in S_1}\left(I_0^W(x) - I_0^M(x)\right)^2},$$

in the case that $0 \leq \alpha_1 \leq 1$.

Proof This is given by a direct computation of the derivative. Let us write

$$\psi(\alpha) = \sum_{x\in S_1}\left(I(x) - \alpha I_0^M(x) - (1-\alpha)I_0^W(x)\right)^2,$$

and note that

$$\psi(\alpha) = \sum_{x\in S_1}\left(I(x) - I_0^W(x) + \alpha(I_0^W(x) - I_0^M(x))\right)^2.$$

Then

$$\begin{aligned}\frac{\partial\,\psi(\alpha)}{\partial\,\alpha} &:= \sum_{x\in S_1} 2\left(I_0^W(x) - I_0^M(x)\right)\left(I(x) - I_0^W(x) + \alpha\left(I_0^W(x) - I_0^M(x)\right)\right. \\ &= 2\sum_{x\in S_1}\left(I_0^W(x) - I_0^M(x)\right)\left(I(x) - I_0^W(x)\right) \\ &+2\,\alpha\sum_{x\in S_1}\left(I_0^W(x) - I_0^M(x)\right)\left(I_0^W(x) - I_0^M(x)\right).\end{aligned}$$

The solution of the equation $\partial\,\psi(\alpha)/\partial\,\alpha = 0$ gives the result.

4 A real case analysis: exporting the ARWU index from a subset of the best universities to a larger set

In this section we address the problem that we introduced in the first section of the document. We worked with some records of ARWU scores that we chose from the top 100 universities along with the records of Incites to test our method. We will explain the procedure step by step.

4.1 Material and methods

After checking several relevant university rankings and databases, we set the problem by using Incites (year 2018) as a source for the variables appearing in the computation of the distance between universities, which provide the similarity relationship. We decided to use the variables "Times Cited" (Total amount of citations of all the papers published by the university in the corresponding year), and the (four) variables given by the number of published papers by quartile ("Articles Q1", "Articles Q2", "Articles Q3", "Articles Q4"). Complementarily, we used the popular Shanghai ranking structure (ARWU ranking, based on the ARWU score) to define the index that we want to check. Specifically, we followed the next steps.

1) We first made an investigation about which records for high level universities could be used for our purpose. The first problem is to identify a set of institutions for which both information sources cited above are clearly presented; it must be taken into account that sometimes classifications are not the same in both databasis. For example, University of California is not univocally defined, since for some databasis different centers belonging to this institution are presented as separated entities. After comparing, we got a maximal set of 84 universities to work with. For them, we were able to obtain the variables that were needed in Incites and the records of the ARWU score.
2) The aim was to use a subset of top universities as a reference for training the model. The way of choosing such a group was to divide the total set of institutions by the ARWU score, taking as training set the upper one. Under the idea of making a 50% division —that is, half for training and half for checking—, we center the corresponding cut-off value of the ARWU score around 30. However, this parameter has been changed for checking the model in the interval $[25, 35]$, what provides a systematic way of changing the size of the training set.
3) Thus, the idea was to use the rest of the universities (the bottom of the ARWU score list) to check the model. As we explained in the previous section, the final extension of the ARWU score for the top set is made by means of a convex combination of the McShane and the Whitney formulae, which gives the corresponding self-defined index. Figure **??** provides a representation for two different training sets, together with the original ARWU score for the best value of α given below (0.69). In the axis OX we represent our 84 universities by their order numbers, which are related to their total size. The labels SDIndex31 and SDIndex33 mean that the training sets are defined for all the universities with an ARWU Score bigger than 31 and 33, respectively. The errors made for the training sets defined in

this way for the values of the ARWU Score 29, 31 and 33 are shown in Figure **??**. Note that we are representing the extended functions, and so the approximations coincide with the original index when the universities belong to the training sets. We follow this criterion in all the figures presented below.

Fig. 2 Representation of the self-defined index training with three different sets, together with the original values of the ARWU score ($\alpha = 0.69$).

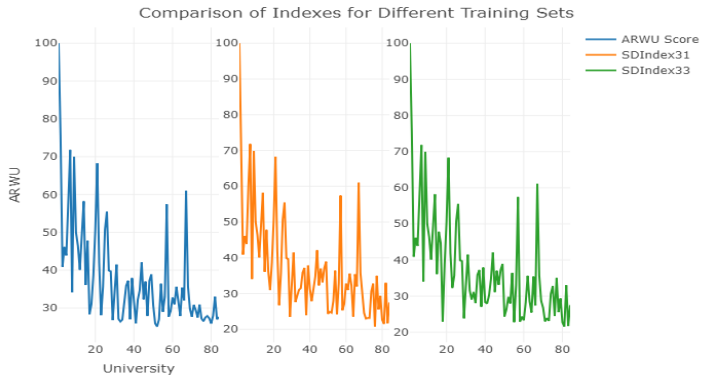
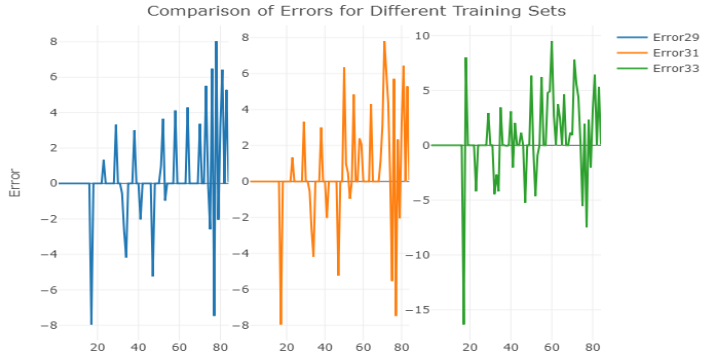


Fig. 3 Representation of the error for three different training sets.



The best parameter α is obtained by means of an optimization method using the error defined in Remark **??**. Taking into account the relative weights that could be given to define the metric in the model, we finally decided to fix it as a weighted Euclidean norm, trying to get a right balance among all variables. The formula is

$$d(u_1,u_2) = \left(\frac{10^{-4}}{64}\cdot\big(\text{Times Cited}(u_1) - \text{Times Cited}(u_2)\big)\right)^2$$

$$+\sum_{i=1}^{4} 2^{i-1}\cdot 10^{-4}\big(\text{Articles } Q_i(u_1)-\text{Articles } Q_i(u_2)\big)^2\Big)^{1/2},$$

where u_1, u_2 belong to the fixed set D of 84 universities. Of course, the decision maker can change these weights according to her/his preferences, or using complementary information that she/he could obtain. Figure **??** provides a representation for three different values of α, together with the original ARWU score. The errors commited are represented in Figure **??**.

Fig. 4 Representation of different solutions for different values of the interpolation parameter $\alpha = a$ together with the original values of the ARWU score.

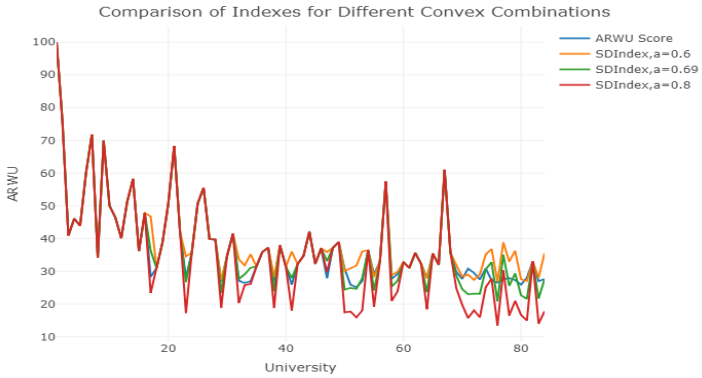
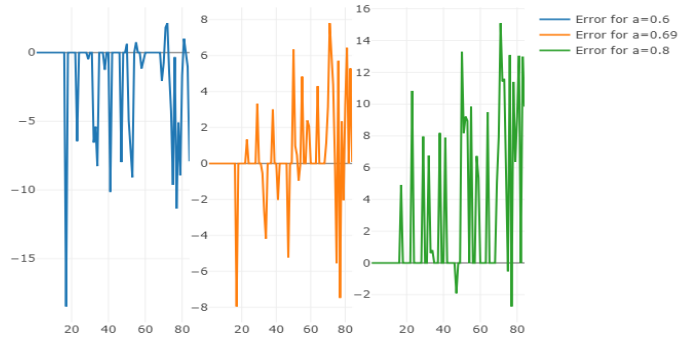


Fig. 5 Representation of the error commited by three different values of $\alpha = a$.



(4) Then we train the algorithm. With the set of universities having ARWU score bigger than 31, we got a Lipschitz constant for the function —we called it the coherence constant Q for the index in the previous sections— equal to 5.826876.

Table 1 Some predicted values of the ARWU Score, together with the error.

TimesCited	Articles Q1	Articles Q2	Articles Q3	Articles Q4	ARWU Score	$I_\alpha^{M,W}$	Error
35562	3849	1581	598	229	29.2	27.1	2.1
32825	3427	1167	526	185	27.9	23.6	4.3
31486	3351	1473	666	266	29.8	28.6	1.2
31128	2982	1219	444	166	27.7	24.5	3.2
30560	3832	1338	647	263	30.8	23	7.8
30302	2988	1265	538	189	29.5	23.1	6.4
30164	3740	1375	530	157	27.5	23.1	4.4
28679	2888	1017	498	164	27.2	32.7	5.5
28178	2959	1228	665	264	26.5	20.7	5.8
25950	2954	1426	510	195	27.9	25.5	2.4

At this point, the algorithm is ready for computing both the McShane and the Whitney extensions. This algorithm can be found in the complementary material.

(5) Finally, the parameter α that optimizes the error —that is, the addition of the squares of the differences among the values of the ARWU score and our self-defined index $I_\alpha^{M,W}$ for the universities with ARWU score lower than 31— is obtained. There are 52 universities in the training set (for ARWU Score ≥ 31), and 32 in the complementary test set. This allows to check the model, by comparing the ARWU score and our extension $I_\alpha^{M,W}$. We show the results in the next section.

The interested reader can find the ready-to-use R algorithm in the Supplementary Material (McShaneWhitneyExt.R).

4.2 Results and discussion

Let us present the results of our experiment for the situation explained above. After trying different training sets, all of them given by the top part of the list and defined using the criterion *ARWU Score* $\geq$ *me* for a given value *me* $\in [0,100]$, we obtained the best result for *me* $= 31$.

As we said above, the best value of the interpolation parameter α obtained by minimizing the error was $\alpha = 0.69$, that is, the final formula for the self defined index is given by

$$I_\alpha^{M,W}(u) = 0.69\, I^M(u) + 0.31\, I^W(u), \quad u \in D.$$

In Table **??**, some predicted values for universities with values of the ARWU scores below 31 are presented, together with the original ARWU Score and the error.

Figures **??** and **??** provide a graphical representation of the best solution for the training set defined by the top universities of our list which have an ARWU Score

bigger to 31, —composed by 52 universities—-, and the error (Figure **??**). As the reader can see, the relative errors commited are reasonable in most of the cases, taking into account that the variables used for defining the metric are not explicitly appearing in the definition of the ARWU Score. Also, it can be seen in Figure **??** that the errors around the last part of the list of universities (that correspond to low values of the ARWU Score) are meaningfully bigger, although they still present an acceptable rate. The total relative error—that is, the addition of the squares of the differences divided by the addition of the square of the values—, is $\varepsilon^2 = 0.0233$.

Fig. 6 Representation of the best solution together with the original values of the ARWU score ($\alpha = 0.69$).

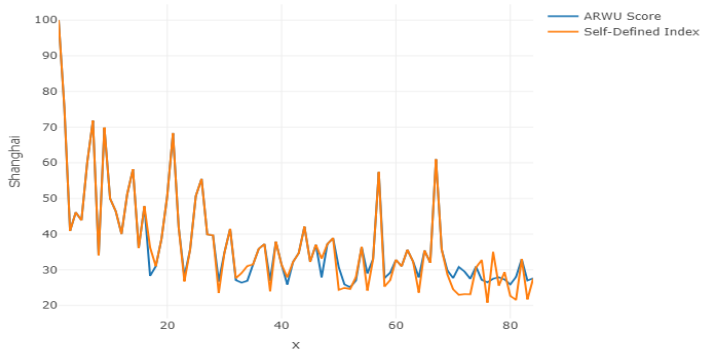
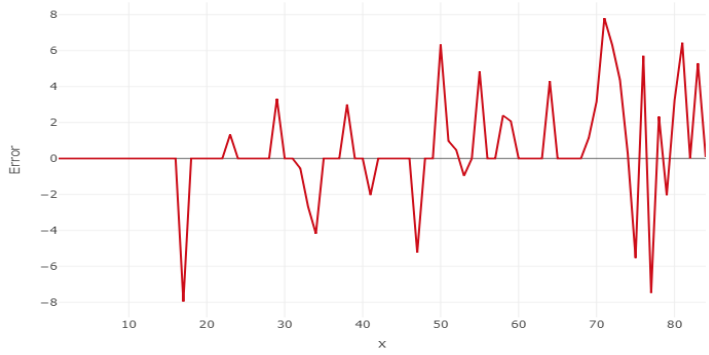


Fig. 7 Representation of the error commited by the best choice of convex combination $\alpha = 0.69$.



Note that only the values of the distances between all the elements of the set of universities are needed, together with the value of the Lipschitz constant (coherence constant of the model), and the original index for the elements of the training

sets. Adding more variables to the selected set for the definition of the metric could improve the results in a meaningful way, even if they are not apparently connected with the definition of the index. Moreover, the way the distance matrix is defined —symmetric matrix composed by all the distances among elements of the underlying metric space—, allows to increase the training set in an easy way when more information is included. In order to do this, it is enough to compute a new column for the matrix, given by all the distance of the new element introduced the other elements of the set. So, the proposed method can be used for defining an iterative self-improving tool, that is, a dynamic system that can be improved continuously under a typical reinforcement learning scheme.

4.3 Conclusions

We have presented a new mathematical structure for extrapolating values of indexes associated to the scientific and educational activity. It is based on the construction of a metric space which represents the similarity relation among items, and the optimization of the convex combination of two extremal extensions of a Lipschitz functions —that represent the index in the model—, that are the McShane and Whitney extensions of Lipschitz functions.

Using our method, we have trained the model to predict some values of the ARWU Score for a subset of top universities using other set of top universities. The trained model could now be used for predicting the values of other universities for which the ARWU Score is not computed, but for which we can find bibliometric values in Incites. To show our technique, we have used the variables TimesCited (number of citations in 2018 to documents published by the university), and the number of published papers in Q1, Q2, Q3 and Q4 the same year.

Although aparently these variables are not directly connected with the ARWU Score, the results fit well, as can be seen in the figures and the data provided in the section of results. Two main conclusions can be stated. First: there is a clear direct conection among successful scientific production and the position in the ARWU list; and second: maybe that the use of sophisticated variables –that are sometimes difficult to measure, or highly restrictive, as having Nobel Prizes— for the definition of university rankings are not really needed.

However, the main conclusion of the paper is the method itself, that provides an easy reinforcement procedure for the extrapolation of indices to sets for which they are not known and cannot be directly computed. For example, it allows to make a prediction of the values of such index for the universities of countries that are not appearing in the ARWU list, but for which Incites (a very big database) has bibliometric records.

Acknowledgements The third and fourth authors gratefully acknowledge the support of the Ministerio de Ciencia, Innovación y Universidades (Spain), Agencia Estatal de Investigación, and FEDER, under grant MTM2016-77054-C2-1-P. The first author gratefully acknowledge the support of Cátedra de Transparencia y Gestión de Datos, Universitat Politècnica de València y Generalitat Valenciana, Spain.

References

1. Aguillo, I., Bar-Ilan, J., Levene, M., & Ortega, J. (2010). Comparing university rankings. *Scientometrics,* 85(1), 243-256.
2. Asadi, K., Dipendra, M., & Littman, M.L. (2018). Lipschitz Continuity in Model-based Reinforcement Learning. Proceedings of the 35th International Conference on Machine Learning, *Proc. Mach. Lear. Res.* 80, 264-273.
3. Bougnol, M.L., & Dulá, J.H. (2013). A mathematical model to optimize decisions to impact multi-attribute rankings. *Scientometrics,* 95(2), 785-796.
4. Cancino, C. A., Merigó, J.M., & Coronado, F.C. (2017). A bibliometric analysis of leading universities in innovation research. *Journal of Innovation & Knowledge,* 2(3), 106-124.
5. Chen, K-H., & Liao, P-Y. (2012). A comparative study on world university rankings: a bibliometric survey. *Scientometrics,* 92(1), 89-103.
6. Cobzaş, Ş., Miculescu, R., and Nicolae, A. (2019). *Lipschitz functions.* Berlin; Springer.
7. Deza, M.M, & Deza, E. (2009). *Encyclopedia of distances.* Berlin: Springer.
8. 2019 U-Multirank ranking: European universities performing well. https://ec.europa.eu/education/news/u-multirank-publishes-sixth-edition-en
9. Cinzia, D., & Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. *Journal of the Association for Information Science and Technology,* 68.2, 508-529.
10. Dobrota, M., Bulajic, M., Bornmann, L., & Jeremic, V. (2016). A new approach to the QS university ranking using the composite I-distance indicator: Uncertainty and sensitivity analyses. *Journal of the Association for Information Science and Technology,* 67(1), 200-211.
11. Falciani, H., Calabuig, J.M., & Sánchez Pérez, E.A. (2020). Dreaming machine learning: Lipschitz extensions for reinforcement learning on financial markets. *Neurocomputing* 398, 172–184.
12. Kehm, B. M. (2014). Global university rankings—Impacts and unintended side effects. *European Journal of Education,* 49(1), 102-112.
13. Lim, M.A., & Øerberg, J.W. (2017). Active instruments: on the use of university rankings in developing national systems of higher education. *Policy Reviews in Higher Education,* 1(1), 91-108.
14. Çakır, M.P., Acartürk, C., Alaşehir, O., & Çilingir, C. (2015). A comparative analysis of global and national university ranking systems. *Scientometrics,* 103(3), 813-848.
15. Luo, F., Sun, A., Erdt, M., Raamkumar, A. S., & Theng, Y. L. (2018). Exploring prestigious citations sourced from top universities in bibliometrics and altmetrics: a case study in the computer science discipline. *Scientometrics,* 114(1), 1-17.
16. von Luxburg, U., & Bousquet, O. (2004). Distance-based classification with Lipschitz functions, *Journal of Machine Learning Research,* 5, 669-695.
17. Marginson, S. (2014). University rankings and social science, *European Journal of Education,* 49(1), 45-59.
18. Pagell, R. A. (2014). Bibliometrics and university research rankings demystified for librarians. In *Library and information sciences* (pp. 137-160). Berlin: Springer.
19. Rao, A. (2015). *Algorithms for Lipschitz Extensions on Graphs,* Yale University: ProQuest Dissertations Publishing, 10010433.
20. Rosa, K. D., Metsis, V., & Athitsos, V. (2012). Boosted ranking models: a unifying framework for ranking predictions, *Knowledge and information systems,* 30(3), 543-568.
21. Saisana, M., d'Hombres, B., & Saltelli, A. (2011). *Rickety numbers: Volatility of university rankings and policy implications.* Research policy, 40(1), 165-177.
22. Tabassum, A., Hasan, M., Ahmed, S., Tasmin, R., Abdullah, D. M., & Musharrat, T. (2017). University ranking prediction system by analyzing influential global performance indicators. In *2017 9th International Conference on Knowledge and Smart Technology (KST)* (pp. 126-131) IEEE.
23. Van Raan, A. F.J., Van Leeuwen, T.N., & Visser, M.S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics,* 88(2), 495-498.