Forum

Editor's Choice

Oikos 126: 1523–1532, 2017
doi: 10.1111/oik.04400

Subject Editor: Christopher Swan. Editor-in-Chief: Dustin Marshall. Accepted 10 April 2017

Avoiding erroneous citations in ecological research: read before you apply

Martin Šigut, Hana Šigutová, Petr Pyszko, Aleš Dolný, Michaela Drozdová and Pavel Drozd

M. Šigut, H. Šigutová (http://orcid.org/0000-0003-1134-248X) (sigutova.hanka@gmail.com), P. Pyszko, A. Dolný, M. Drozdová and P. Drozd, Dept of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, Univ. of Ostrava, Chittussiho 10, 710 00 Ostrava, Czech Republic.

The Shannon–Wiener index is a popular nonparametric metric widely used in ecological research as a measure of species diversity. We used the Web of Science database to examine cases where papers published from 1990 to 2015 mislabelled this index. We provide detailed insights into causes potentially affecting use of the wrong name 'Weaver' instead of the correct 'Wiener'. Basic science serves as a fundamental information source for applied research, so we emphasize the effect of the type of research (applied or basic) on the incidence of the error. Biological research, especially applied studies, increasingly uses indices, even though some researchers have strongly criticized their use. Applied research papers had a higher frequency of the wrong index name than did basic research papers. The mislabeling frequency decreased in both categories over the 25-year period, although the decrease lagged in applied research. Moreover, the index use and mistake proportion differed by region and authors' countries of origin. Our study also provides insight into citation culture, and results suggest that almost 50% of authors have not actually read their cited sources. Applied research scientists in particular should be more cautious during manuscript preparation, carefully select sources from basic research, and read theoretical background articles before they apply the theories to their research. Moreover, theoretical ecologists should liaise with applied researchers and present their research for the broader scientific community. Researchers should point out known, often-repeated errors and phenomena not only in specialized books and journals but also in widely used and fundamental literature.

Synthesis

Basic science serves as a fundamental information source for applied research, and these should work as one unit. The Shannon index is an ecological parameter that is used by both applied and fundamental ecologists but we identified systematic differences in the erroneous usage of this metric between these two groups. The index is more popular among applied research scientists but they also more likely to erroneously mislabel the index, despite the error in mislabeling being pointed out years ago. Incorrect citations are a persistent problem in ecology but we emphasize the problematic communication between basic and applied science: explorations into citation culture revealed that almost 50% of authors have not actually read their cited.

Reference lists as well as the Introduction and Discussion sections of published papers provide valuable sources for readers in search of specific information (McLellan et al. 1992). However, Simkin and Roychowdhury (2003a) estimated that only about 20% of authors read the original papers, despite the fact that copying a citation without reading it can lead to "cloning misprints" when the secondary source is incorrect, create citation errors (Nicolaisen 2007), produce renowned papers (Simkin and Roychowdhury 2003b), and worst of all, lead to an incorrect interpretation of the original ideas. Large-scale copying of incorrect references has been documented several times and has been considered a serious problem (Ball 2002, Clarke 2003, Todd and Ladle 2008). An exemplary case comes from quantitative ecology and the Shannon–Wiener diversity index. The name of the second author, Norbert Wiener, is often replaced by Warren Weaver, despite multiple warnings published not only in scientific papers (Spellerberg and Fedor 2003) but also in well-known ecological textbooks (Krebs 1999, Magurran 2004).

The Shannon–Wiener index is one of the most common nonparametric indices, originally proposed by Claude Shannon (considered a founder of information theory, Verdú 1998) to quantify entropy in information (Shannon 1948), and based on the work of Norbert Wiener, known as a father of informatics (Conway and Siegelman 2006). The 'Shannon–Wiener measure' has appeared in scientific papers since 1950. Good (1953) used the index itself for the first time in ecology as a measure of animal population heterogeneity. While Good (1953) cited Shannon's original paper (Shannon 1948), MacArthur (1955), often considered the

first to introduce the Shannon–Wiener index into ecology, cited a book by Shannon and Weaver (1949) that contained Shannon's original work from 1948. The book included a second part titled "Recent contribution to the mathematical theory of communication," written by mathematician Warren Weaver (Spellerberg and Fedor 2003). In fact, it seems that MacArthur himself was the first to mislabel the index. The combination of the Shannon–Wiener index with the citation of Shannon and Weaver (1949) probably led some authors to think that "Wiener" (Smith and Wilson 1996, sometimes incorrectly spelled as Weiner; Samuelson 2001) was a typographical error of "Weaver" (Spellerberg and Fedor 2003). Ironically, in the abstract of his part of the book, Weaver himself quoted Claude Shannon's proclamation that information theory owed a great debt to Norbert Wiener (Shannon and Weaver 1949), and readers of the original text would do well to note this fact. On the other hand, Weaver also mentioned Shannon's formula in his part of the text, which could lead many authors to believe that he is likewise an author of the entropy measurement. Either way, Berg (1979) was probably the first to point out the incorrect attribution of the index name to Weaver, and many subsequent publications have concurred, notably popular methodological textbooks of Krebs (1999) and Magurran (2004), and Spellerberg and Fedor (2003) even explicitly recommended labelling the "Shannon–Wiener index" in the title of their paper.

Names from unfamiliar languages are often typed incorrectly and errors in the name of the first author are common (Kotiaho et al. 1999, Buchanan 2006, Sweetland 1989). However, the mislabeled Shannon–Wiener diversity index is different and probably stems from confusion regarding names and inattentive copying without consulting the original source. In general, the most problematic citation errors arise from seminal works or papers that are cited more frequently than others (Garfield 1972), as any potential mistake spreads apace. The probability that a paper will be cited is influenced by many factors (Taborsky 2009), which can create citation bias, leading to many forms of errors (Garfield 1990, Nicolaisen 2007, Bornmann and Daniel 2008). Despite the many studies dealing with the effects of a paper's research focus and target journal on citation errors (Goldberg et al. 1993, Hansen and McIntire 1994, Lok et al. 2001), authors have not considered the type of research (basic versus applied). Basic research serves as a source of information for applied research (Rosenberg 1990, Roll-Hansen 2009), and many authors have emphasized the positive results of their integration (Munoz-Sanjuan and Bates 2011). Therefore, we hypothesized that applied science has incorporated concepts from basic science without a thorough consideration of the theoretical background (primary sources), and subsequently, has not been as quick to include newer findings, thereby producing a higher frequency of citation errors in papers focused on applied research.

We investigated mislabeling of the Shannon–Wiener diversity index using the Web of Science database. Specifically, we aimed to 1) examine the trends in using the incorrect name 'Shannon–Weaver' instead of the correct 'Shannon–Wiener' in biological research, and 2) investigate patterns of use of the incorrect name (with a special focus on the type of research).

Material and methods

Trends in (mis)labeling the index

We used the Web of Science (WOS) database (Web of Science Core Collection, KCI Korean Journal Database, SCIelo Citation Index and Russian Science Citation Index) (Thomson Reuters 2016) to find the frequency of correct and incorrect forms of the index name, using appropriate built-in WOS basic search filters (under 'topic', from 1990–2015, refined by 'document types: article'). We used the search string *"Shannon Wiener" OR "Shannon Weaver" NOT ("Shannon Weaver communication" OR "Shannon Weaver information")* to find all papers mentioning only diversity indices and to exclude papers mentioning problems of the Shannon–Weaver communication model or information theory (Shannon 1948, Shannon and Weaver 1949). To find papers using both alternatives (some authors propose using both terms or conversely advise against using both), we used *"Shannon Wiener" AND "Shannon Weaver"*. Then we used *"Shannon Weaver" NOT ("Shannon Weaver communication" OR "Shannon Weaver information") NOT ("Shannon Weaver" AND "Shannon Wiener")* to find only papers using the incorrect index name, omitting those using both alternatives. We exported all records mentioning either correct (Shannon–Wiener, S–Wi) or incorrect names (S–We) and noted the paper's and journal's title, first author's address, publication year, and research area using 'marked list' filters in the WOS database. We assigned country of origin to each paper according to the first author's address and determined the region according to country divisions in SCImago (2007). If addresses were not included in the database, we searched on addresses or other information within particular paper or in an Internet search engine. Then we went through the exported records and excluded duplicates and papers referring to information or communication theory and conference presentations that were not eliminated via search strings or filters. We compared the proportion of papers using the incorrect name Shannon–Weaver (S–We) to those using any form of 'Shannon' index (S–W). Further, we compared the proportion of papers mentioning a Shannon index (S–W) to all relevant biological papers over the years. We obtained numbers of biological papers within the years and regions by advanced searching of paper records comprising at least one of the top ten biological research areas (which were stated according to their prevalence within all papers mentioning a S–W index) using the following general string: *SU = (area 1 OR area 2 OR … OR area 10) AND DT = Article AND AD = (country 1 OR country 2 OR … OR country n) AND PY = (year)*.

We analyzed trends in the use of the S–W index and the mislabeling frequency in R (<www.r-project.org>). We performed arcsine transformations of all proportional data to normalize them, and then converted them to a 'time series object' format. To assess changes in use of the index (S–W/biological papers) and mislabeling frequency (S–We/S–W), we used a trend test based on a nonparametric Spearman test between the observations and time (R package 'pastecs'). We subsequently adjusted final p-values using a false discovery rate correction (Siegel and Castellan 1988).

Patterns of mislabeling

To analyze patterns of mislabeling, we chose 1) the focus of the paper (research area, journal, basic versus applied research), 2) the first author's affiliation (country and world region), 3) the interaction of both 1) and 2), and 4) the willingness of authors to read cited literature (primary citations).

Papers in the WOS database can be classified in several research areas, so we went through all abstracts (if available) to strictly categorize them as applied (practical) or basic (theoretical), according to the Organisation for Economic Co-operation and Development definition (2002). From analysis comparing applied and basic science, we excluded papers that did not fit either category.

We used weighted analysis of variance to investigate trends in the use of the Shannon index in biological research (S–W/biological papers) and the rate of error (S–We/S–W) over time in 1) basic and applied research categories and 2) individual regions (Chambers et al. 1992). Based on the results of 2), we split regions a posteriori in two categories, with higher or lower impact on the scientific community as reflected by the total H-index in SCImago. The impact of countries in individual regions may not be evenly distributed, so we subsequently tested trends in the use of the Shannon index in biological research (S–W/biological papers) and the frequency of mislabeling (S–We/S–W) among countries according to their H-index (countries with a total H-index value in SCImago >200 versus the rest of the world). For testing trends in proportion of papers using Shannon index in biological research, we used analysis weighted by the total number of biological papers. For testing trends in the rate of error we used analysis weighted by the number of papers using Shannon index. The models were originally built from two basic variables: specific explanatory variable (research, region or H-index) and year, and moreover from the polynomial function of the second degree for year and the interaction term. Backward selection was then used to reduce model in the case of insignificant interaction or polynomials, however, because the interactions were focal for our hypothesis, their significance is always mentioned, regardless of further reduction of the model. In the case that the residuals of the model were not normally distributed, we performed arcsine transformation to normalize dependent variable. The start of the monitored timeline 1990–2015 was set to zero to enhance the intuitiveness of the parameter estimates interpretation. We compared the difference in the mislabeling frequency in English and non-English-speaking countries using a χ^2 test.

We assessed the willingness of authors to read cited sources by calculating the proportion of papers presenting the incorrect name while also citing the primary source with the correct name. We searched for the citations of the original Shannon (1948) or Shannon and Weaver (1949) papers or their reprints, as well as other crucial ecological works dealing correctly with diversity indices (Krebs 1999, Magurran 2004, Pielou 1975) in WOS using the Cited reference search engine. Then we combined results from Cited reference search engine with search results of all records with the incorrect name (S–We, obtained by above mentioned formula) in search history.

Data deposition

Original data from WOS database are available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.9m808> (Šigut et al. 2017).

Results

Trends in (mis)labeling the index

In total, we found 2098 scientific papers presenting the index (S–Wi or S–We, i.e. S–W); of these, 661 (31.5%) used the incorrect name (S–We), 1435 (68.4%) used the correct name (S–Wi), and 2 (0.1%) used both names concurrently. The proportion of all papers using the S–W index (S–W/biological papers) increased over time ($df=24$, $\rho=0.958$, $p<0.001$), while the proportion of papers using the incorrect name (S–We/S–W) decreased over time ($df=24$, $\rho=-0.813$, $p<0.001$).

Patterns of mislabeling

Focus of the paper (research area, journal, basic versus applied research)

In a comparison of research areas, Environmental Sciences and Ecology (687), Marine and Freshwater Biology (378), Agriculture (307), Plant Sciences (210), and Biodiversity and Conservation (152) mentioned the Shannon index (S–W) the most often, however, when adjusting for the area size (S–W/total number of biological papers per research area), the highest preference for the index was in Oceanography (7.2×10^{-04}), Marine and Freshwater Biology (7.0×10^{-04}), Forestry (4.3×10^{-04}), Biodiversity and Conservation (3.1×10^{-04}) and Environmental Sciences and Ecology (2.8×10^{-04}). Areas that are primarily applied research had the highest proportion of papers using incorrect name (S–We/S–W) (Table 1). Similarly, applied research journals had the highest proportion of mistakes (S–We/S–W) (Table 2).

Of the 2098 scientific papers mentioning the S–W index, we considered 1332 (63.5%) to be applied research, 759 (36.2%) to be basic research, and 7 (0.3%) to be unsuitable for further analysis owing to their ambiguous nature. We assigned 875 (61.2%) and 554 (38.8%) of the 1429 papers with the correct name (S–Wi) to the applied and basic research categories, respectively. We assigned 457 (69.2%) and 203 (30.8%) of the 660 papers presenting the incorrect name (S–We) to the applied and basic research categories, respectively.

The proportion of papers using the Shannon index (S–W/biological papers) was higher in applied research than in basic ($df=46$, $F=187.6$, $p<0.001$); it has been generally increasing for both research categories together ($df=46$, $F=162.0$, $p<0.001$) but with significant interaction term signalizing different shape of trends ($df=46$, $F=37.9$, $p<0.001$) with constantly increasing trend in applied research but stagnant trend in basic research in recent years (Fig 1a).

The slope of the trend (the interaction) in the proportion of papers using the incorrect name to all papers mentioning the Shannon index (S–We/S–W) did not significantly differ between categories ($df=47$, $F=1.53$, $p=0.223$). After

Table 1. Research areas with the highest proportion of papers mentioning the incorrect name of the index to the total number of papers mentioning the Shannon index (Shannon–Weaver/Shannon total) and their assignment to research categories based on a proportion of applied papers (basic: 0–25%, basic/applied: 26–50%, applied/basic: 51–75%, applied: 76–100%; included only areas with at least 20 Shannon–Weaver papers within research area; every paper could have been assigned to more than one research area).

	Research area	Shannon total	S–Weaver/S total	Applied/S total	Research category
1.	Agronomy	131	0.618	0.916	applied
2.	Genetics and Heredity	36	0.611	0.694	applied/basic
3.	Soil Science	95	0.579	0.937	applied
4.	Agriculture	307	0.537	0.928	applied
5.	Microbiology	122	0.525	0.672	applied/basic
6.	Biotechnology and Appl. Microbiology	92	0.511	0.924	applied
7.	Plant Sciences	210	0.448	0.695	applied/basic
8.	Engineering	69	0.406	0.971	applied
9.	Water Resources	62	0.323	0.919	applied
10.	Life Sciences and Biomedicine	110	0.264	0.373	basic/applied

model simplification the trend has been decreasing over time in both research categories ($df = 49$, $F = 64.8$, $p < 0.001$; Fig. 1b). However, we found a significantly higher proportion of mislabeling within the applied research category ($df = 49$, $F = 13.2$, $p < 0.001$; Fig. 2).

First author's affiliation

The comparison of individual regions and countries by first author's affiliation showed that the highest number of papers mentioning the S–W index originated from Asia (n = 610, 29.1%; of this, China = 58.3%), western Europe (n = 464, 22.1%; of this, Spain = 16.8%), Latin America (n = 379, 18.1%; of this, Brazil = 48.8%) and northern America (n = 280, 13.3%; of this, USA = 79.3%), while the lowest numbers originated from eastern Europe (n = 140, 6.67%; of this, Poland = 43.9%), Africa (n = 90, 4.3%; of this, Tunisia = 18.9%), the Middle East (n = 83, 3.96%; of this, Turkey = 60.2%), and Pacific region (n = 52, 2.48%; of this, Australia = 65.4%). However, the highest preference of S–W index (S–W/total number of biological papers per region) showed Latin America (8.0×10^{-04}), followed by Africa (4.9×10^{-04}), Asia (3.7×10^{-04}), eastern Europe (3.6×10^{-04}) and Middle East (2.9×10^{-04}), while the lowest was in western Europe (1.7×10^{-04}), Pacific region (1.5×10^{-04}) and northern America (1.0×10^{-04}).

The proportion of papers mentioning the S–W index to the total number of relevant biological papers (S–W/biological papers) varied among individual regions ($df = 180$, $F = 73.2$, $p < 0.001$), and trends over time differed significantly ($df = 180$, $F = 20.8$, $p < 0.001$; Fig. 3).

The proportion of papers mislabeling the index to the total number of papers mentioning the Shannon index (S–We/S–W) showed a similar trend in all regions ($df = 178$, $F = 0.91$, $p = 0.498$; Fig. 4) but after further reduction of the model differed among individual regions ($df = 173$, $F = 3.19$, $p < 0.01$).

By region, the proportion of papers with the S–W index was significantly higher in those whose impact on the scientific community was lower (Africa, the Asiatic region,

Table 2. Journals with the highest proportion of papers mentioning the incorrect name of the index to the total number of papers mentioning the Shannon index (Shannon–Weaver/Shannon total) and their assignment to research categories based on a proportion of applied papers (basic: 0–25%, basic/applied: 26–50%, applied/basic: 51–75%, applied: 76–100%; journals focusing predominantly on basic research are in bold). Only journals with at least five Shannon–Weaver results are included. IF is impact factor.

	Journal	Shannon total	S–Weaver/S total	Applied/S total	IF [2015]	Research category
1.	Euphytica	8	1.000	0.750	1.618	applied
2.	Hereditas	7	1.000	0.714	1.118	applied/basic
3.	Philippine Agricultural Scientist	5	1.000	1.000	0.266	applied
4.	Genetic Res. and Crop Evolution	36	0.944	1.000	1.258	applied
5.	Biology and Fertility of Soils	10	0.900	1.000	3.069	applied
6.	Crop Science	8	0.875	1.000	1.575	applied
7.	Pedosphere	6	0.833	0.833	1.500	applied
8.	Revista Arvore	15	0.667	0.467	0.296	**basic/applied**
9.	Fems Microbiology Ecology	14	0.643	0.571	3.530	applied/basic
10.	Bioresource Technology	8	0.625	1.000	4.917	applied
11.	Applied and Env. Microbiology	13	0.615	1.000	5.932	applied
12.	Canadian J. of Microbiology	10	0.600	0.700	1.335	applied/basic
13.	Soil Biology and Biochemistry	10	0.600	1.000	4.152	applied
14.	J. of Environmental Biology	17	0.588	0.353	0.563	**basic/applied**
15.	Applied Soil Ecology	21	0.571	1.000	2.670	applied
16.	Microbial Ecology	17	0.412	0.529	3.232	applied/basic
17.	Polish Journal of Ecology	20	0.350	0.450	0.500	**basic/applied**
18.	Hydrobiologia	44	0.295	0.295	2.051	**basic/applied**
19.	Revista de Biologia Tropical	21	0.286	0.381	0.524	**basic/applied**
20.	Estuarine Coastal and Shelf Science	19	0.263	0.316	2.335	**basic/applied**

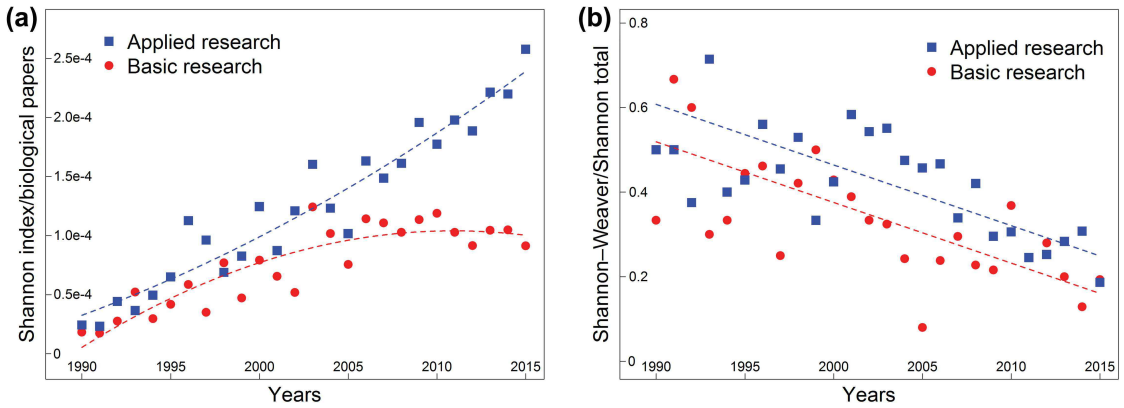

Figure 1. (a) Proportion of papers mentioning the Shannon index to the total number of biological papers (Shannon total/biological papers). Equations for model parameter estimates are for applied research: $\text{proportion}_{(\text{S–W/biological papers})} = 1.250 \times 10^{-04} + 4.468 \times 10^{-04} \times \text{year}_{1990} + 3.905 \times 10^{-05} \times \text{year}_{1990}^2$, and for basic research: $\text{proportion}_{(\text{S–W/biological papers})} = 7.529 \times 10^{-05} + 2.052 \times 10^{-04}\ \text{year}_{1990} - 8.145 \times 10^{-05} \times \text{year}_{1990}^2$. (b) Proportion of papers using the incorrect name of the index to the total number of papers mentioning the Shannon index (Shannon–Weaver/Shannon total) according to research category (applied or basic) in individual years from 1991 to 2015. Equations for model parameter estimates are for applied research: $\text{proportion}_{(\text{S–We/S–W})} = 0.607620 - 0.014342 \times \text{year}_{1990}$, and for basic research: $\text{proportion}_{(\text{S–We/S–W})} = 0.518995 - 0.014342 \times \text{year}_{1990}$.

eastern Europe, Latin America, the Middle East) than where the impact is considered higher (northern America, western Europe, the Pacific region) ($df = 192$, $F = 244.8$, $p < 0.001$). The proportion sharply increased after 2000 in the former regions, but it was relatively unchanged in the latter ($df = 192$, $F = 64.2$, $p < 0.001$). The differences in the frequency of mislabeling (S–We/S–W) and in trends between these two regional groups were also significant ($df = 178$, $F = 5.03$, $p = 0.026$; $df = 178$, $F = 4.60$, $p = 0.033$, respectively).

The highest proportion of mislabeling (S–We/S–W) occurred within Africa (0.444), followed by eastern Europe (0.400), northern America (0.368), the Middle East (0.361), Asia (0.305) and western Europe (0.298), while the lowest proportions were present within the Pacific region (0.255) and Latin America (0.251).

There was a higher preference for the S–W index (S–W/biological papers) in the countries with higher H-index values (> 200) when compared to the rest of the world ($df = 48$, $F = 884$, $p < 0.001$). However, there was a difference in mislabeling frequency (S–We/S–W) between both groups, with higher rate of error in the latter ($df = 47$, $F = 5.99$, $p = 0.018$), with decreasing trend in both groups ($df = 47$, $F = 59.9$, $p < 0.001$) but with no interaction between trends ($df = 45$, $F = 1.85$, $p = 0.181$). The highest proportion of mislabeling (countries with at least 10 search results of all papers mentioning the Shannon index) occurred within Russia (0.778), followed by Ethiopia (0.769), the Philippines (0.636), and Argentina (0.552), while the lowest proportions were present within Chile (0.050), Belgium (0.067), Norway (0.105), and Mexico (0.108). A higher proportion of mislabeling (S–We/S–W) originated from native English-speaking countries (0.337) than from non-English-speaking countries (0.310), but it was not significantly different ($\chi^2 = 0.895$, $df = 1$, $p = 0.344$).

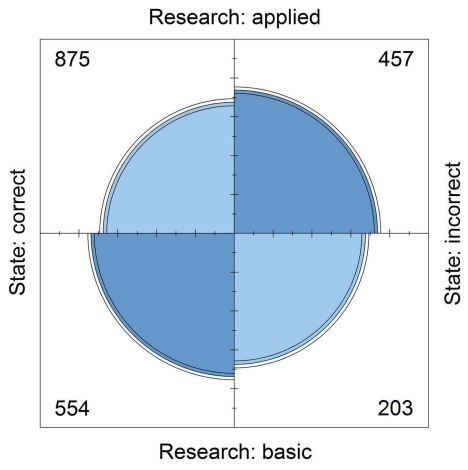

Figure 2. A fourfold plot (standardized by preserving odd ratios) showing a significantly higher proportion of papers using the incorrect name of the index (Shannon–Weaver) within applied research.

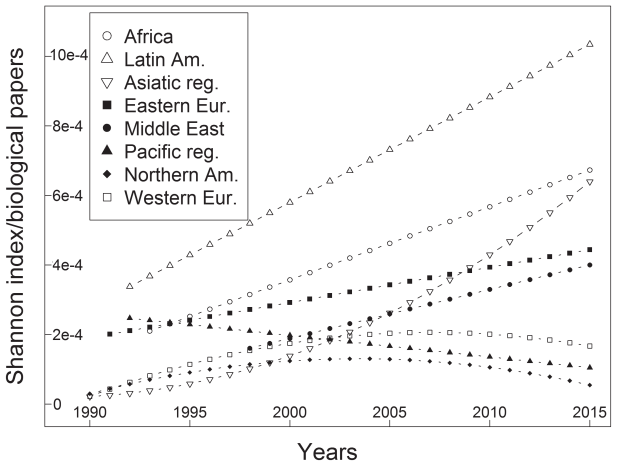

Figure 3. Trends in using the Shannon–Wiener diversity index over time in individual regions, given by the number of papers mentioning any form of the Shannon index to the total number of relevant biological papers published from 1990 to 2015.

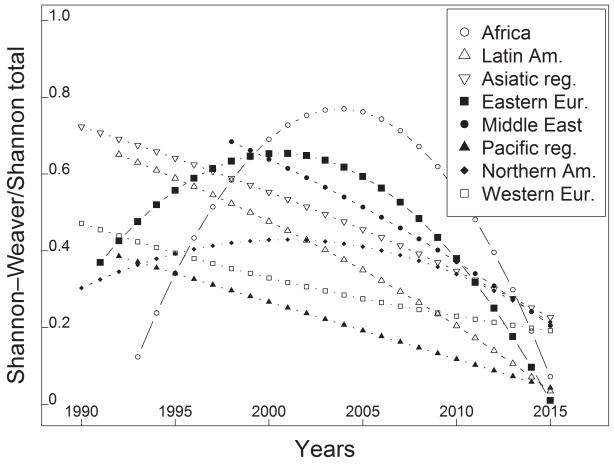

Figure 4. Trends in mislabeling of the Shannon–Wiener index in individual regions over time. The graph shows the number of papers mentioning the incorrect name of the Shannon index to the total number of papers mentioning the Shannon index, published from 1990 to 2015.

Combination of authors' affiliation and focus of the paper

The highest proportion of papers that mentioned the S–W index and focused on applied research occurred within the Asiatic region (0.737) followed by northern America (0.709), Africa (0.674), western Europe (0.616), Pacific region (0.596), and eastern Europe (0.579), while the lowest proportion occurred within Latin America (0.505) and the Middle East (0.458). Of countries with at least 10 papers that mentioned the S–W index and focused on applied research, the Philippines had the highest proportion (1.000, n = 11), followed by Ethiopia (0.923, n = 12), Portugal (0.816, n = 31), Belgium (0.800, n = 12), and China (0.777, n = 276), while Colombia (0.200, n = 4), Tunisia (0.353, n = 6), Turkey (0.360, n = 18), Chile (0.421, n = 8), and Mexico (0.446, n = 33) had the lowest proportions.

The highest proportions of papers (n) that mislabeled the index (S–We) in applied research were in Africa (where research that incorrectly stated the name of the index was 4.7-fold more prevalent in applied than basic papers, n = 40), followed by the Asiatic region, northern America, Latin America, western Europe, Pacific region and eastern Europe (n = 186, 3.5-fold; n = 102, 2.9-fold; n = 95, 2-fold; n = 138, 1.9-fold; n = 13, 1.2-fold; n = 56, 1.2-fold, respectively), while only in the Middle East was there a higher proportion of papers with mistakes originating from basic research (n = 30, 1.5-fold). Comparing individual countries (with at least five papers mentioning the S–We index), the proportion of applied research papers mentioning the incorrect (S–We) name was highest in the Philippines, where all of the papers with an incorrectly stated name were focused on applied research (n = 7), followed by Ethiopia, Israel, China and the United Kingdom (n = 10, 9-fold more papers originated from applied research; n = 7, 6-fold; n = 86, 5.6-fold; n = 13, 5.5-fold, respectively). The proportion was lower in Turkey, Mexico, Tunisia, Colombia and Spain, which had fewer papers with mistakes in the basic research category (n = 16, 15-fold; n = 8, 1.7-fold; n = 5, 1.5-fold; n = 5, 1.5-fold; and n = 22, 1.4-fold, respectively).

The willingness of authors to read the cited sources

Of the 661 papers that incorrectly named the index (S–We), 323 (48.9%) had cited a primary reference that provided the correct name. The remaining papers (338, 51.1%) 1) cited a source that provided the incorrect name, 2) cited a rarely cited source that provided the correct name (but that should be probably considered a secondary source), or 3) did not provide any citation for the index (Table 3).

Discussion

Trends in (mis)labeling the index

Although many renowned textbooks and papers have summarized problems of diversity indices or provided critical reviews of their use (Hurlbert 1971, Šustek 1980, Krebs 1999, Magurran 2004), indices adopted from information theory are still considered appropriate, their use over time has been increasing, and, as our study showed, the Shannon–Wiener index in particular has become even more popular. Why? There might be one simple reason: scientists favor a mathematical method that produces a number.

We found that about one-third of all scientific papers cited in WOS used an incorrect name for this index. While relatively high, this proportion is decreasing, probably due to the increasing number of publications pointing out this particular error (Krebs 1999, Spellerberg and Fedor 2003,

Table 3. The references predominantly cited as a primary source of the used index and the number of citations within 661 papers with the incorrect name of the index (Shannon–Weaver); Web of Science database (December 2016).

	Reference	Source of index name	No. of citations	No. of citations (WOS)
1.	Shannon and Weaver (1949)	correct (original)	146	11 266
2.	Shannon and Weaver (1963)	correct (original reprint)	56	2743
3.	Magurran (1988)	correct	44	5928
4.	Pielou (1975)	correct	24	2310
5.	Shannon (1948)	correct (original)	17	21 249
6.	Margalef (1957)	correct	16	1023
7.	Magurran (2004)	correct	6	3562
8.	Good (1953)	correct	5	1511
9.	Krebs (1999)	correct	4	2743
10.	MacArthur (1955)	correct	3	940
11.	Spellerberg and Fedor (2003)	correct	2	162
	other	incorrect, secondary, or missing citation	338	N/A

Magurran 2004) or highlighting the problems of citations in general (Simkin and Roychowdhury 2003a, Todd and Ladle 2008, Taborsky 2009). A growing number of communication channels, particularly on the Internet, is facilitating clarification of possible misunderstandings among authors (especially via social networks: e.g. Research Gate, Facebook and Twitter scientific groups), as is better accessibility to primary literature due to the increasing number of scientific databases. Nevertheless, this aspect could be a double-edged sword, either reducing citation errors or biases by detecting and avoiding them, or conversely, creating confusing heterogeneity and leading to their increase. Improving journal policy due to the pressure for publishing lower number of errors and fewer biases (Garfield 1990), as well as increasing use of citation tools and managers (e.g. freeware Zotero), electronic reference databases and editing services, or obviously increasing number of co-authored papers (Plume and van Wiejen 2014) could also play crucial roles in avoiding errors before the manuscript is submitted. However, careful checking by authors, reviewers, and editors during manuscript submission and review prevents most errors (Lok et al. 2001).

Patterns of mislabeling

First author's affiliation

Authors' affiliations are a crucial factor that influences citation errors. Authors from different countries or world regions have to deal with language barriers, different politics, research funding, or access to the literature. English-speaking countries produce at about 50% of the world's papers and have a dominant position in the scientific world (May 1997). Standard journal policies often require non-native speakers to write articles in English (Bakewell 1992, Tregenza 2002), which may produce more citation errors due to the language barrier (Jacobs et al. 2006, Anonymous 1981). Therefore, incorrect citations should be more common with authors from non-English-speaking countries (Kotiaho et al. 1999). Contrary to our expectations, we found a slightly higher (although not significant) proportion of mistakes in papers originating from English-speaking countries, suggesting that native speakers may be less conscientious during manuscript preparation, or non-native speakers might seek professional editing which may reveal errors before the manuscript submission. However, it should be pointed out that an author's affiliation address does not necessarily indicate his/her origin or native language.

Another influence connected to an author's affiliation is the access to scientific resources among countries or regions (Møller 1990, May 1997, King 2004), which is largely dictated by economic differences. Not all journals are equally accessible to scientists from different countries, as many universities have limited subscriptions to electronic journals; authors therefore may only be able to read an abstract, which increases the probability of making a mistake. Regions with better access thus should have fewer citation biases and errors (Todd and Ladle 2008). Moreover, in certain countries, researchers might be required to publish in national journals (Mishra 2008) that might have lower quality standards and lax citation policies.

Although we found differences in mislabeling among regions with different access, contrary to our expectations, we found a relatively high proportion of mislabeling within Northern America and a relatively low proportion within Latin America, which might reflect a native language difference (English- versus non-English-speaking countries). Similarly, we found a high incidence of mislabeling from the Russian Federation. Nevertheless, the expected trend was clear in the comparison of individual countries, with the lowest proportion of mislabeling from Belgium and Norway but surprisingly also from Chile and Mexico. This inconsistency may stem from conditions prevailing in particular countries or regions, or may be just from a generally lower incidence of papers with the S–W index in countries with lower H-index values; our results thus should be treated with caution.

The proportion of papers in which the authors used the Shannon–Wiener index was higher in regions with less access. Authors from Africa, Asia and Latin America had the highest increase in its use. The highest proportion of papers using the Shannon index focused on Environmental Sciences and Ecology, so the higher incidence may reflect an increasing scientific effort in ecological and environmental fields within these regions. Developing countries may have a greater interest in economic growth than in minimizing environmental damage, when compared to developed economies that addressed environmental issues decades ago (Hart and Cavanagh 2012). However, these regions are increasingly judged not solely on their economic progress but instead must balance progress with environmental protection. This situation has recently changed in many countries owing to the adoption of environmental policies and management practices. For instance, India made great progress from 1995 to 2010 in addressing its environmental issues and improving environmental quality (World Bank 2011). Similar efforts are also documented in China (Zhiyong 2004) and Latin America (Jenkins 2000), and may be extrapolated to all countries that have recently started to be more involved in their environmental issues. In conclusion, a combination of growing research on the environment, rapid economic and publication growth, less access to sources pointing out the limitations of diversity indices, or maybe a preference for indices is probably responsible for a higher preference for Shannon–Wiener diversity index in certain regions.

Focus of the paper (research area, journal, basic vs applied research)

According to Roll-Hansen (2009), the distinction between basic and applied research categories can be dependent on the subjective decision of a particular researcher, and individual papers often include both aspects and do not strictly fit either category. However, we read complete abstracts and only excluded seven (0.3%) papers from subsequent analysis because we could not classify them.

We found a higher proportion of papers presenting the incorrect name within applied research, probably stemming from the different nature of the types of research. Basic is defined as experimental or theoretical work performed to obtain new knowledge without any particular application or use, while applied is primarily directed to practical use (Organisation for Economic Co-operation and Development 2002). For example, in her ecological textbook, Magurran (2004) warned against incorrect attribution of the index name to Warren Weaver. We believe that an ecologist

studying functional diversity and community structure of microorganisms in soils and a biologist dealing with bioremediation of contaminated soil using a microbial community have different likelihoods of having read this book. It is surprising that to date no study has considered the type of research as an aspect potentially influencing the occurrence of citation errors.

The highest proportion of mislabeling occurred within journals dealing mostly with applied research, supporting our hypothesis of more frequent citation errors within this type of research. Only six of 20 journals with the highest proportion of mislabeling predominantly published basic research papers. According to Lok et al. (2001), the incidence of citation errors may be also affected by the scientific quality of journals, as authors publishing in lower-quality journals may be less careful while preparing their reference list. Despite strong criticism of the impact factor concept (Alberts 2013, Jacobs 2009, PLoS Medicine Editors 2006), it is still recognized as a relevant measure (Olden 2007) as it can reflect the quality of a particular journal (Saha et al. 2003). The six journals indicated above have relatively low impact factors (mean = 1.10, median = 0.56, SD = 0.93), and their high rate of erroneous labeling of the index might reflect this fact. Moreover, rules set by journals with lower impact factors may be generally less strict, and they might be more indulgent with authors originating from a particular country. Our results support this hypothesis as a high rate of erroneous labeling of the index occurred in Polish Journal of Ecology; although it is focused mainly on basic research, most of the publishing authors (15 of 20 papers) were native Poles. Despite the fact that the highest number of papers mentioning the S–W index belonged to the area Environmental Sciences and Ecology and dealt mainly with basic research, the highest proportion of mistakes occurred within areas primarily focused on applied research. This was the case for Agronomy and Agriculture, the research areas primarily directed to practical use, which together had one of the highest numbers of papers mentioning the index and subsequently also the highest proportion of mislabeling.

Although the proportion of mislabeling has decreased over time in both research categories, there was no significant difference in the decline between categories, but the shape of the curve showed a lag for applied research, supporting our hypothesis that it incorporates new knowledge from basic science relatively slowly. In the future, this delay could result in a situation when biologists focused on basic research will cease using the incorrect name of the index, while in applied research, this error will continue to spread for many years. Also, due to the increased steering of research priorities, commercialization, and broader accountability of science, applied research is growing faster at the expense of basic research (Nowotny et al. 2003), and thus may be more susceptible to the occurrence of citation errors. Therefore, researchers focused on applied research should be more careful when selecting sources during manuscript preparation.

Interaction between authors' affiliation and focus of the paper

The prevalence of research types in a certain country or region depends on the character of its practical problems, opportunities, general social needs, existing economy, and political goals (Roll-Hansen 2009). This is consistent with our results, as we found different proportions of papers mentioning the Shannon index and focused on applied research within individual regions and countries. Considering the fact that a higher proportion of mistakes was found within applied research, we suggest that the authors' affiliation in combination with the type of research may play a substantial role in the frequency of erroneous labeling of the index. Most strikingly, Chinese authors showed one of the strongest preferences for the Shannon–Wiener diversity index, and these papers mostly focused on applied research (0.777). Taking into account that in Chinese papers focused on applied research, the frequency of mislabeling was about 69.8 % higher than in papers focused on basic research, the total effect of this country on the frequency of citation error was probably more significant than that of many other countries.

The willingness of authors to read the cited sources

Our results suggest that almost 50% of the authors presenting the incorrect name in their papers (i.e. 15.4% from a total of 2098 papers mentioning any form of Shannon index) probably had not read the cited sources. This proportion is high even when compared to the estimation of Simkin and Roychowdhury (2003a), which was much higher; only 20% of authors had read the original source. However, the proportion of authors only using secondary citations was probably even higher. We calculated results based only on the most cited seminal sources among papers that incorrectly stated the name of the index, so we expect that a substantial number of remaining sources also provide the correct name. Authors using the correct name might also have drawn from secondary sources, further increasing this enormous proportion. Our results support the hypothesis of Spellerberg and Fedor (2003), who suggested that the key factor in mislabeling the Shannon–Wiener diversity index was quoting and requoting references without going back to the original sources.

Conclusions

Despite the strong criticism since Hurlbert (1971), species diversity indices seem to be an immortal phenomenon and are common "community metrics" in many ecological, environmental, or agricultural studies. This is also true for the Shannon–Wiener diversity index, which is still used, and seems to be ever more popular, especially in applied disciplines such as biotechnology, agriculture, biochemistry, and biomedicine. Our study indicates persistent problems in citation culture. The Shannon–Wiener diversity index is commonly mislabeled, especially in applied disciplines, which probably stems from using secondary sources without consulting the primary literature. Scientists dealing with applied research thus should take care to select and consult the basic research literature for their discipline before they apply methods to their own research. Authors' willingness to read the primary literature is a reflection of whether the communication of basic and applied science works in the first place. To improve this communication, theoretical ecologists should raise public awareness of frequent errors by including mention of them in general, widely used, and popular textbooks as well as in specialized literature.

Acknowledgements – Editage (<www.editage.com>) provided professional editing of the manuscript.
Funding – This project was funded by the Ministry of Education, Youth and Sports of the Czech Republic (no. LO1208), the Institute of Environmental Technologies (no. CZ.1.05/ 2.1.00/03.0100), the Grant Agency of the Czech Republic (GA14-04258S), and the Grant Agency of University of Ostrava (no. SGS20/PrF/2015). Financial support through M. Roeselová Memorial Fellowship is gratefully appreciated.
Author contribution statement – PD, AD, MD, MŠ and HŠ designed the research. PP and MŠ performed the statistical analyses. MŠ, HŠ and PD wrote the manuscript.

References

Alberts, B. 2013. Impact factor distortions. – Science 340: 787.
Anonymous 1981. What is a surname? – Essays of an Information Scientist 82: 26–30.
Bakewell, D. 1992. French research – publish in English, or perish. – Nature 356: 648.
Ball, P. 2002. Paper trail reveals references go unread by citing authors. – Nature 420: 594.
Berg, J. 1979. Discussion of methods of investigating the food of fishes, with reference to a preliminary study of the prey of *Gobiusculus flavescens* (Gobiidae). – Mar. Biol. 50: 263–273.
Bornmann, L. and Daniel, H.-D. 2008. What do citation counts measure? A review of studies on citing behavior. – J. Doc. 64: 45–80.
Buchanan, R. A. 2006. Accuracy of cited references: the role of citation databases. – Coll. Res. Libr. 67: 292–303.
Chambers, J. M. et al. 1992. Analysis of variance; designed experiments. – In: Chambers, J. M. and Hastie, T. J. (eds), Statistical models in S. Wadsworth and Brooks, pp. 145–193.
Clarke, T. 2003. Copied citations give impact factors a boost. – Nature 423: 373.
Conway, F. and Siegelman, J. 2006. Dark hero of the information age: in search of Norbert Wiener, the father of cybernetics. – Basic Books.
Garfield, E. 1972. Citation analysis as a tool in journal evaluation. – Science 178: 471–479.
Garfield, E. 1990. Journal editors awaken to the impact of citation errors. How we control them at ISI. – Essays of an Information Scientist 41: 33–11.
Goldberg, R. et al. 1993. Reference accuracy in the emergency-medicine literature. – Ann. Emerg. Med. 22: 1450–1454.
Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. – Biometrika 40: 237–264.
Hansen, M. E. and McIntire, D. D. 1994. Reference citations in radiology: accuracy and appropriateness of use in two major journals. – Am. J. Roentgenol. 163: 719–723.
Hart, M. and Cavanagh, J. 2012. Environmental standards give the United States an edge over China. – Center for American Progress. <www.americanprogress.org/issues/green/news/2012/04/20/11503/environmental-standards-give-the-united-states-an-edge-over-china/>.
Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. – Ecology 52: 577–586.
Jacobs, D. et al. 2006. What do third world researchers lack? Documenting the peer review. – Curr. Sci. 91: 1605–1607.
Jacobs, H. 2009. Pay to cite. – EMBO reports 10: 1067.
Jenkins, R. 2000. Industry and environment in Latin America. – Psychology Press.
King, D. A. 2004. The scientific impact of nations. – Nature 430: 311–316.
Kotiaho, J. S. et al. 1999. Unfamiliar citations breed mistakes. – Nature 400: 307.
Krebs, C. J. 1999. Ecological methodology. – Benjamin/Cummings.
Lok, C. K. et al. 2001. Risk factors for citation errors in peer-reviewed nursing journals. – J. Adv. Nurs. 34: 223–229.
MacArthur, R. 1955. Fluctuations of animal populations and a measure of community stability. – Ecology 36: 533.
Magurran, A. E. 1988. Ecological diversity and its measurement. – Princeton Univ. Press.
Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.
Margalef, R. 1957. Information theory in ecology. – Gen. Syst. Bull. 3: 36–71.
May, R. M. 1997. The scientific wealth of nations. – Science 275: 793–796.
McLellan, M. F. et al. 1992. Trust, but verify. The accuracy of references in four anesthesia journals. – Anesthesiology 77: 185–188.
Mishra, D. C. 2008. Citations: rankings weigh against developing nations. – Nature 451: 244.
Møller, A. 1990. National citations. – Nature 348: 480.
Munoz-Sanjuan, I. and Bates, G. P. 2011. The importance of integrating basic and clinical research toward the development of new therapies for Huntington disease. – J. Clin. Invest. 121: 476–483.
Nicolaisen, J. 2007. Citation analysis. – Ann. Rev. Inf. Sci. Technol. 41: 609–641.
Nowotny, H. et al. 2003. "Introduction. Mode 2" revisited: the new production of knowledge. – Minerva 41: 179–194.
Olden, J. D. 2007. How do ecological journals stack-up? Ranking of scientific quality according to the h index. – Ecoscience 14: 370–376.
Organisation for Economic Co-operation and Development 2002. Frascati manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development: the Measurement of Scientific and Technological Activities. – OECD Publications.
Pielou, E. C. 1975. Ecological diversity. – Wiley.
PLoS Medicine Editors 2006. The impact factor game. – PLoS Med. 3: e291.
Plume, A. and van Wiejen, D. 2014. Publish or perish? The rise of the fractional author. – Res. Trends 38. – <www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/>.
Roll-Hansen, N. 2009. Why the distinction between basic (theoretical) and applied (practical) research is important in the politics of science. – London School of Economics and Political Science, Contingency and Dissent in Science Project.
Rosenberg, N. 1990. Why do firms do basic research (with their own money)? – Res. Policy 19: 165–174.
Saha, S. et al. 2003. Impact factor: a valid measure of journal quality? – J. Med. Libr. Assoc. 91: 42.
Samuelson, G. M. 2001. Polychaetes as indicators of environmental disturbance on subarctic tidal flats, Iqaluit, Baffin Island, Nunavut Territory. – Mar. Pollut. Bull. 42: 733–741.
SCImago 2007. SJR – SCImago journal and country Rank. – <www.scimagojr.com>.
Shannon, C. E. 1948. A mathematical theory of communication. – Bell Syst. Tech. J. 27: 379–423.
Shannon, C. E. and Weaver, W. 1949. The mathematical theory of communication. – Univ. Illinois Press.
Shannon, C. E. and Weaver, W. 1963. The mathematical theory of communication. – Univ. Illinois Press (reprint).
Siegel, S. and Castellan, J., N. Jr. 1988. Nonparametric statistics for the behavioral sciences. – McGraw-Hill.
Šigut, M. et al. 2017. Data from: Avoiding erroneous citations in ecological research: read before you apply. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.9m808>.
Simkin, M. V. and Roychowdhury, V. P. 2003a. Read before you cite! – Complex Systems 14: 269–274.

Simkin, M. V. and Roychowdhury, V. P. 2003b. Copied citations create renowned papers? – Ann. Improbable Res. 11: 24–27.
Smith, B. and Wilson, J. B. 1996. A consumer's guide to evenness indices. – Oikos 76: 70–82.
Spellerberg, I. F. and Fedor, P. J. 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon–Wiener" Index. – Global Ecol. Biogeogr. 12: 177–179.
Šustek, Z. 1980. Použitie Shannon–Wienerovej funkcie k posudzovaniu narušenia ekosystémov. – Lesnický výskum a výchova vedeckých pracovnikov v ČSSR.
Sweetland, J. 1989. Errors in bibliographic citations – a continuing problem. – Libr. Q. 59: 291–304.
Taborsky, M. 2009. Biased citation practice and taxonomic parochialism. – Ethology 115: 105–111.
Thomson Reuters 2016. Web of Science [v.5.22] – <https://webofknowledge.com>.
Todd, P. A. and Ladle, R. J. 2008. Hidden dangers of a "citation culture." – Ethics Sci. Environ. Politics 8: 13–16.
Tregenza, T. 2002. Gender bias in the refereeing process? – Trends Ecol. Evol. 17: 349–350.
Verdú, S. 1998. Fifty years of Shannon theory. – IEEE Trans. Inf. Theory 44: 2057–2078.
World Bank 2011. Environmental assesment. Country data: India. – World Bank Group.
Zhiyong, L. 2004. A policy review on watershed protection and poverty alleviation by the Grain for Green Program in China. – In: Sim, H. C. et al. (eds), Forests for poverty reduction: opportunities with clean development mechanism, environmental services and biodiversity. Food and Agriculture Organization of the United Nations Regional Office for Asia and the Pacific, pp. 133–138.