



Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? An information-theoretical and statistical estimation approach

Rüdiger Mutz¹ 

Received: 29 October 2021 / Accepted: 25 February 2022 / Published online: 19 March 2022
© The Author(s) 2022

Abstract

Diversity is a central concept not only in ecology, but also in the social sciences and in bibliometrics. The discussion about an adequate measure of diversity is strongly driven by the work of Rao (Sankhyā Indian J Stat Series A 44:1-22, 1982) and Stirling (J R Soc Interface 4:707-719, 2007). It is to the credit of Leydesdorff (Scientometr 116:2113-2121, 2018) to have proposed a decisive improvement with regard to an inconsistency in the Rao-Sterling-diversity indicator that Rousseau (Scientometr 116:645-653, 2018) had pointed out. With recourse to Shannon's probabilistically based entropy concept, in this contribution the three components of diversity "variety", "balance", and "disparity" are to be reconceptualized as entropy masses that add up to an overall diversity indicator div_e . Diversity can thus be interpreted as the degree of uncertainty or unpredictability. For "disparity", for example, the concept of *mutual information* is used. However, probabilities must be estimated statistically. A basic estimation strategy (cross tables) and a more sophisticated one (parametric statistical model) are presented. This overall probability-theoretical based concept is applied exemplarily to data on research output types of funded research projects in UK that were the subject of the Metric Tide Report (REF 2014) and ex-ante evaluation data of a research funding organization. As expected, research output types depend on the research area, with journal articles having the strongest individual balance among the output types, i.e., being represented in almost all research areas. For the ex-ante evaluation data of 1,221 funded projects the diversity components were statistically estimated. The overall diversity of the projects in terms of entropy is 55.5% of the maximal possible entropy.

Keywords Interdisciplinarity · Diversity · Entropy · Mutual information · Multinomial distribution

✉ Rüdiger Mutz
ruediger.mutz@uzh.ch

¹ Center of Higher Education and Science Studies, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland

Introduction

Diversity is not only a central concept in scientometrics (e.g., interdisciplinarity), but also in ecology (e.g., Jost, 2007) or in the social sciences (e.g., Brusco et al., 2020; Jones & Dovidio, 2018). Therefore, the measurement and operationalization of this concept is of eminent importance, which is also reflected in numerous publications on this concept especially in the field of ecology (see Foredean et al. (2017)). A certain break in the development of this concept comes from the work of Rao and Stirling (Rao, 1982; Stirling, 2007), who developed a diversity concept based on the existing literature, which, in brief, consists of three components:

- *Variety* «Variety is the number of categories into which system elements apportioned. It is the answer to the question: “How many types of things we have? ... All else being equal, the greater the variety, the greater the diversity.» (Stirling, 2007, p. 709),
- *Balance* «Balance is a function of the pattern of apportionment of elements across categories... All else being equal the more even is the balance, the greater the diversity.» (Stirling, 2007, p. 709),
- *Disparity* “Disparity refers to the manner and degree in which the elements may be distinguished... All else being equal, the more disparate are the represented elements, the greater the diversity.” (Stirling, 2007, p. 709).

Thus, first Rao (1982) and later Stirling (2007, p. 721) developed a "general diversity heuristic" as a sum indicator of different elements j, i of a system (e.g., different research area in research proposals):

$$D = \sum_{j(i \neq j)} d_{ij} p_i p_j \quad (1)$$

where p_i and p_j are proportions of elements i and j in the system (as base for balance), and d_{ij} is the degree of difference between the elements (disparity). Variety is the number of different elements. This concept had a strong influence not only in ecology (e.g., Rousseau et al., 1999) but also in scientometrics (e.g., Goyanes et al., 2020; Rousseau, 2018; Wang, Thijs, & Glänzel, 2015). The idea of composing an indicator multiplicatively from several individual indicators is captivating and opens up many possibilities for analysis. It is also easy to calculate.

Unfortunately, Rousseau (2018, p. 651) was able to show with a simple data example that a central assumption of the diversity indicator ("monotonicity") formulated by Stirling (2007, p. 711) does not hold. For a given variety and disparity the measure does not increase monotonically with balance. This points to fundamental problems with this concept. It is to the credit of Leydesdorff (2018) and Leydesdorff, Wagner, and Bornmann (2019b) to create an indicator that does not have these problems and continues to multiplicatively link the three individual diversity indicators, which is certainly a viable approach despite criticism and modifications (Leydesdorff et al., 2019a; Rousseau, 2019).

For $i = 1$ to n_c and $j = 1$ to n_c categories, the diversity is defined as follows (Leydesdorff, 2018, p. 2116):

$$Div_c = (n_c/N) \cdot GINI \cdot \left[\sum_{\substack{i=1 \\ j=1 \\ i \neq j}}^{i=n_c \\ j=n_c} \frac{d_{ij}}{(n_c \cdot (n_c - 1))} \right], \tag{2}$$

where n_c is the number of categories, N the total number of categories, GINI is the Gini-coefficient, d_{ij} is the disparity the Euclidean distance in terms of 1-cosine (Ahlgren et al., 2003), which is normalized by the factor $(n_c \cdot (n_c - 1))$. Unlike the classical Euclidean measure, d_{ij} in terms of 1-cosine varies between 0 and 1. The Bravais-Pearson correlation is not suitable for measuring similarity taking values between -1 and $+1$. Additionally, the cosine measure is invariant for adding zeros.

In this paper, a different path shall be taken, namely a path back to the information-theoretical roots of this concept, which is based on probabilistic foundations as already used by Shannon (1948). Although the initial work of Shannon and the entropy concept is mentioned in most publications and is also part of Stirling’s concept, the probabilistic foundation and the classical entropy concept seems to have been lost or is currently not sufficiently received. For example, in a paper in Scientometrics, Shaw Jr. (1981) derives his concept of information theory from a thermodynamic representation of entropy, in which probabilities are only implicitly used. But why is probability theory for a consistent diversity concept so important? Four main reasons can be given:

1. *Unclear theoretical basis:* The theoretical basis for combining the different elements of diversity is not immediately apparent. For example, probabilities as p_j or p_i are linked to correlations or similarity measures (d_{ij} , Eq. 1, 2), which is not statistically derived from any probabilistic or statistical framework. Even Rao (1982, p. 7) himself says about the choice of d_{ij} : “The choice of d_{ij} is not a statistical problem and will depend on an individual’s assessment of differences between qualitative categories reference to a given problem. However, one can use methods of multidimensional scaling in estimating d_{ij} by using *supplementary information* such as inequality relationships between d_{ij} and d_{rs} for different (i, j) and (r, s) .”
2. *Accuracy:* Most diversity indicators are defined numerically, although the relative frequencies or probabilities are statistical quantities and thus have a statistical measurement error. Probabilities based on small sample sizes are far less accurate than probabilities estimated from large samples.
3. *Interpretation of indicators:* Given the large number of diversity indicators, the question remains as to how to interpret the individual indicators and what advantages one indicator has over another indicator.
4. *Statistical analysis:* Diversity indicators often have to be processed statistically, which in turn requires distributional assumptions. Therefore, the question is whether the data should not be formulated in a probabilistic form in order to apply the statistical analysis more or less directly to the raw data.

The central idea of the paper is to refer back to Shannon’s concept of entropy (Leydesdorff & Ivanova, 2021), which is based on probability theory and forms the basis of today’s diversity concepts, to arrive at a modified concept of diversity that firstly avoids previous inconsistencies in the diversity concept, secondly allows a simple form of interpretation in terms of uncertainty and reduction of uncertainty (entropy), and thirdly forms the basis both for new types of diversity indicators and for statistical modelling. When it comes to

an application, the probabilities must be determined numerically, and that's when statistics comes into play. Probabilities are not fixed quantities but, like any statistical parameter, have estimation errors that are higher the smaller the sample size. Additionally, statistical models make it possible to include explanatory variables to explain, predict or adjust probabilities. Therefore, the information-theoretical approach is complemented by a statistical estimation approach. In the following, first the modified diversity concept and an indicator will be derived in probabilistic terms, and second a statistical approach will be outlined how to estimate these diversity components. The approach will be illustrated both with a practical application on research output data from Metrics Tide (<https://responsiblemetrics.org/the-metric-tide/>) and ex-ante evaluation data of projects funded by a Swiss research funding organization. The paper is an extended version (statistical approach, ex-ante evaluation data) of a paper presented at the “International Conference on Scientometrics and Informetrics” (ISSI 2021) (Mutz, 2021).

Methodological approach

Probabilistic framework of information theory and modified diversity indicator div_e

Within information theory, information is defined purely syntactically without reference to semantics based on probability theory. In the following, a simple example will be taken as a starting point: The diversity of research output types of funded projects in the field of Natural Science (e.g., Mutz et al., 2012, 2014). Let us assume that project reports from 1,000 funded projects are available as part of an ex-post evaluation (synthetic data, Table 1) with two outputs. The marginal frequencies of $N=4$ different output types (output 1) are shown in Table 2, on one side before the data analysis without any information (prior or expectations) and on the other side after the empirical data analysis (posterior).

We choose a step-by-step approach. First, we explain the basic concepts of Shannon's information theory using the marginal frequency table of document types (Table 2, output 1) on the assumption that a project produces only one output of a particular type at a time. However, research projects usually produce several outputs, which might be combined, e.g., output 1 "article" and output 2 "article". Therefore, in a second step, it is necessary to open the box by analyzing the combination of document types of the first versus the second output in a cross table “output 1 × output 2” (Table 1). A further information-theoretical

Table 1 Toy example: Cross table of output 1 and output 2 for 1000 funded projects (case II dependence)

Output 1	Output 2				Marginal frequencies
	Books	Articles	Proceedings papers	Reports	
Books	50	0	50	0	100
Articles	0	800	0	0	800
Proceedings papers	50	0	50	0	100
Reports	0	0	0	0	0
Marginal frequencies	100	800	100	0	1000

For instance, 50 projects have proceeding papers and books as outputs, 800 projects have articles as outputs

Table 2 Marginal relative frequencies for different output types (data example) and decision tree

Events	Output type	Code	prior	posterior
1	Books	00	0.25	0.10
2	Articles	10	0.25	0.80
3	Proceedings paper	01	0.25	0.10
4	Reports	11	0.25	0
Entropy			2	0.92

prior= expectation before the data analysis, posterior=probabilities calculated by dividing the marginal frequencies in Table 1 by the total 1000

concept will be introduced, the concept of "mutual information". A very good introduction can also be found in the works of Leydesdorff (Leydesdorff, 1991, 1995; Leydesdorff & Ivanova, 2021).

Step 1: Basic concepts

The following metaphor can be used to define information (Amann & Müller-Herold, 2011, p. 1f). If one wants to know the type of one of the four project outputs, exactly 2 binary questions are necessary to determine the type (Table 2, decision tree). With 8 output types 3 binary questions and with $N=2^k$ types k binary questions are necessary. The information content as number of binary questions is defined as $I=k=\log_2(N)$, if all events are equally likely to occur (prior, Table 2) or no empirical information is available. The maximum information content in the case of equal probabilities is $I=\log_2(4)=2$ bits (see Table 2, code). Another definition of information is obtained from probability theory, where the output type is a random variable X with $j=1$ to N possible events or occurrences (books, articles,...). Thus, the information content is $\log_2(1/p_j)$ or $-\log_2(p_j)$ with $p_j=1/N$ if all events have the same probability (assumption $\log_2(p_j=0)=0$). The rarer the event, the lower the probability, the higher the information. The average information over all events is then:

$$H(X) = \sum_{j=1}^N -p_j \log_2(p_j) \tag{3}$$

and is called Shannon entropy with the assumptions that $\sum_{i=1}^N p_j = 1.0$ and $\log_2(p_j=0)=0$. It is assumed in a first step that the probability of occurrence of one event does not depend on the probability of occurrence of another event, an assumption which will be given up later. In research projects, the publication of proceedings papers should not depend on whether or not a journal article has also been published.

Shannon Entropy, as a measure of information, ultimately expresses the degree of uncertainty. For example, in a coin toss the uncertainty is highest when the probability of heads or tails is 0.5. A probability of 0.8 for heads, for example, reduces the uncertainty

(tampered coin toss). People who bet on "heads" have a greater chance of winning than those who bet on "tails". In the above example, maximum entropy ($H(X)=2$) is reached when there is equal probability ($p_j=0.25$) of the events. However, the actual observed frequencies as an estimate of the probabilities are not equal, the resulting entropy and thus uncertainty is reduced by 54% from 2 to 0.92 bits (see Table 2). The total entropy is the sum of $-0.1*\log_2(0.1) - 0.8*-\log_2(0.8) - 0.1*\log_2(0.1) - 0*\log_2(0) = 0.92$. These considerations are in line with the idea of Bayesian statistics, in which probability is defined as uncertainty. Bayesian Inference can be defined as a statistical learning process in which an initial uncertainty about a parameter defined as a prior probability is reduced in light of the data (posterior probability). (Kruschke, 2011, p. 56f). With the entropy concept, a first measure of diversity can be derived, that is "balance". For example, the more the observed probabilities resemble an equal probability, the higher is the so-called balance, the higher is the entropy or uncertainty.

Shannon also refers to "variety": "With equally likely events there is more choice, or uncertainty, when there are more possible events." (Shannon, 1948, p. 10). Furthermore, variety could also be traced back to an entropy measure. The maximum possible variety would be N , i.e., the maximum possible number of events in a population (here $N=4$ output types). Finally, the maximum variety can be defined as the maximum entropy of the system as follows:

$$H_{Variety_{max}} = \sum_{j=1}^N 1/N \log_2(1/N) = -\log_2(1/N) \tag{4}$$

In the case of four output types, the maximum variety would be $-\log_2(1/4)$. The observed variety corresponds to the number of events with nonzero probability $p_j > 0$:

$$H_{Variety_{obs}} = - \sum_{j=1}^N \left((p_j > 0) \frac{1}{\sum_{j=1}^N (p_j > 0)} \log_2 \left(\frac{1}{\sum_{j=1}^N (p_j > 0)} \right) \right) = -\log_2(1/ \sum_{j=1}^N (p_j > 0)), \tag{5}$$

where $\sum_{j=1}^N (p_j > 0)$ is the number of units (e.g., disciplines) in use. Eventually, $(p_j > 0)$ is a 0/1-indicator variable with 1 (if $p_j > 0$) and 0 (if $p_j = 0$)).

For the above example, the variety_{obs} is $-\log_2(1/3)$. Similarly, (Leydesdorff 2018, p. 2115) argues when defining a relative variety $\sum_{j=1}^N (p_j > 0)/N$. For an common diversity indicator one could add the two entropy masses balance and diversity, although this sum does not represent a joint-distribution $H(A, B)$ in probabilistic terms, but can still be interpreted in terms of entropy with balance and variety as independent quantities:

$$H_{diversity} = H_{variety} + H_{balance} \tag{6}$$

Step 2: "Mutual information"

To get to the last component of diversity, "disparity", the assumption of independent events has to be abandoned. For example, research outputs in a research project may be published in different output types and this may create stochastic dependencies. For example, the

Table 3 Case I: Independence

	Event	Output type	X (Output 2)				Total
			Books	Articles	Proceedings paper	Reports	
Y (Output 1)	1	Books	.01	.08	.01	.00	.10
	2	Articles	.08	.64	.08	.00	.80
	3	Proceedings paper	.01	.08	.01	.00	.10
	4	Reports	.00	.00	.00	.00	.00
	Total		.10	.80	.10	.00	1.00

Table 4 Case II: Dependence

	Event	Output type	X (Output 2)				Total
			Books	Articles	Proceedings paper	Reports	
Y (Output 1)	1	Books	.05	.00	.05	.00	.10
	2	Articles	.00	.80	.00	.00	.80
	3	Proceedings paper	.05	.00	.05	.00	.10
	4	Reports	.00	.00	.00	.00	.00
	Total		.10	.80	.10	.00	1.00

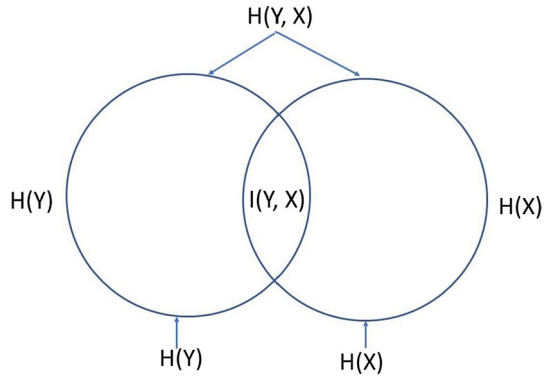
probability of a research article may depend on whether or not a proceedings paper has been published. For simplification, it is assumed that results of a research project are published in two research outputs with a maximum of two different output types, where it is also possible that none of the output types is published or the same output type can be chosen twice. These two research outputs can be defined as two random variables *X* and *Y*.

Table 3 presents a cross-tabulation of cell frequencies and marginal frequencies of the two outputs, where the marginal frequencies of the type of the two outputs are the same as in Table 2, but do not correspond to the cell frequencies of Table 1. It quickly becomes apparent that output 2 does not depend on output 1; the ratios of the relative frequencies across the columns remain the same across the rows. For example, for “articles” the row frequencies divided by the respective marginal frequency remains equal across columns: $0.08/0.10 = 0.64/0.80 = 0.08/0.10 = 0.80$. Thus $p(X|Y) = P(X)$ or $H(X, Y) = H(X) + H(Y)$, the marginal frequencies are sufficient (“overall stochastic balance”).

For Table 4, which corresponds to the data of Table 1, the situation is different: $P(X|Y) \neq P(X)$, i.e., stochastic dependence is present. For example, articles in both outputs occur more frequently ($p = 0.80$) than expected in the case of independence of *X* and *Y* ($p = 0.80 * 0.80 = 0.64$). Two events are stochastically independent, if the probability of *X* AND *Y* equals the product of the probability of *X* and the probability of *Y*: $P(X \cap Y) = P(X) P(Y)$.

In that case, Shannon coined the term *mutual information* $I(X, Y)$, i.e. the reduction of the uncertainty of one random variable by considering another random variable. In the case of two random variables, the mutual information is (Eshima, 2020, p. 8f):

Fig. 1 Entropy measures



$$I(X, Y) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y|X) \tag{7}$$

$$H(X|Y) = H(X) - I(X, Y)$$

The joint entropy of X and Y is then (see Fig. 1):

$$H(X, Y) = H(X) + H(Y) - I(X, Y) \tag{8}$$

The mutual information $I(X, Y)$ can be calculated from the data using the following equation:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right), \tag{9}$$

where $p(x, y)$ is the joint probability of X and Y . If the events are independent, then $p(x, y) = p(x)p(y)$ and $I(X, Y) = 0$. The more the event Y depends on X , the higher the mutual information $I(X, Y)$, the lower the total entropy $H(X, Y)$. Mutual information represents the stochastic dependence of X and Y and is not a correlation. Regarding binary communications Pregowska, Szczepanski, and Wajnryb (2015, p. 1) conclude: “Our research shows that the mutual information cannot be replaced by sheer correlations.” With respect to Table 4 the mutual information for output 1 “article” and output 2 “article” is $I(X=2, Y=2) = 0.80 \log_2(0.80/(0.80 \cdot 0.80)) = 0.80 \log_2(1.25) = 0.258$.

In Fig. 2 the relationships of the different concepts of entropy and mutual information are shown as a Venn diagram (Fig. 1).

Finally, mutual information, already mentioned by Shannon (1948, p. 12), defines the third component of diversity, “disparity”, i.e. the degree of dependence of categories or events. The higher $I(X, Y)$, the lower the disparity. If $I(X, Y)$ is zero, then the disparity is maximal.

In the above example (Table 4), the mutual information is $I(X, Y) = 0.72$, the unconditional entropy of X is $H(X) = 0.92$, and the conditional entropy is $H(X|Y) = 0.20$. This reduces the uncertainty, i.e., overall diversity, by $(0.92 - 0.20)/0.92 = 78.3\%$ when taking into account the dependence in the data in the sense of disparity. Another measure of dependency is Cramer’s V a correlation coefficient, for categorical variables. Comparable to the mutual information $I(X, Y)$, Cramer’s V of 0.29 shows that output 1 and output 2 are correlated.

		Discipline B		
		0	1	Marg. freq.
Discipline A	0	p_{00}	p_{01}	$p_{00}+p_{01}$
	1	p_{10}	p_{11}	$p_{10}+p_{11}$
	Marg. freq.	$p_{00}+p_{10}$	$p_{01}+p_{11}$	$p_{00}+p_{01}+p_{10}+p_{11}=1.0$

ID	A	B	k	d_1	d_2	d_3	d_4
1	0	0	1	1	0	0	0
...
...	1	0	2	0	1	0	0
n_i
...	0	1	3	0	0	1	0
...
...	1	1	4	0	0	0	1
N

Fig. 2 Cross-tabulate and data for the statistical estimation of probabilities for two disciplines: Cross-tabulate with cell frequencies (above), which can be estimated by the data design (below), where d_1 - d_4 are the dummy-codes cells (k) of the cross-tabulate (k =cell number)

In the case of 3 research outputs X, Y and Z , for example, the above formula, corresponding to the addition theorem of probabilities, expands as follows:

$$H(Y, X, Z) = H(Y) + H(X) + H(Z) - I(Y, X) - I(Y, Z) - I(X, Z) + I(Y, X, Z), \quad (10)$$

where the last term of Eq. 10 is positive and represents a correction term in a quantity algebraic representation (i.e., subtraction of an intersection).

Unlike Stirling and Leydesdorff (Leydesdorff, 2018; Leydesdorff et al., 2019a; Stirling, 2007), the different components of the overall diversity indicator div_e are now additive:

$$div_e = H_{variety} + H_{balance} - I_{disparity}, \quad (11)$$

where $div_e = 4.15$ for the above example.

In summary: On the probabilistic basis of Shannon’s entropy concept, a modified diversity concept was derived that is consistent, additive in nature and whose components can be empirically derived. The entropy concept gives the diversity concept a clear interpretation: the higher the diversity, the higher the information, the higher the uncertainty, the less structure there is in the data.

Statistical approach

The concept of diversity as entropy leaves unanswered how the probabilities for calculating entropy can be estimated statistically. To simply determine them by relative frequencies falls short for several reasons. bibliometric data fluctuate randomly and systematically over time because bibliographic data bases are continuously updated (journals, articles, citations, ...). Furthermore, the data bases are incomplete with respect to fields and document types, for instance. Relative frequencies are only good estimates of probabilities when the

sample size goes to infinity. In the case of low sample size or cell frequencies the estimation errors must be taken into account. Statistical models allow to include covariates to explain, predict or adjust probabilities (Bornmann et al., 2021, p. 9317f). Different approaches to the statistical estimation of entropy can be found in the diversity literature, especially in ecology.

In the following, the thematic focus is no longer on the diversity of output types, but on the interdisciplinary nature of projects—the content changes, the methodological concept remains the same (“disciplines” are the “outputs”). In order to conduct statistical analyses, a concrete design for the survey of interdisciplinarity is required, which will guide the statistical analyses. For instance, in the realm of the ex-ante evaluation of project proposals in funding organizations, the simplest way to assess interdisciplinarity is for the principal investigators to list the individual disciplines that are relevant to the planned project. The specified single disciplines can be further grouped into major disciplines at different levels of aggregation. Since several individual disciplines can be assigned to one main discipline, in addition to a dichotomous representation (e.g., main discipline A present or not), which is favored here, a representation can be used that distinguishes different weights of main disciplines in a project proposal (e.g., three individual disciplines from the Humanities (weight=0.75), one individual discipline from the Natural Sciences (weight=0.25)). A simple data analysis strategy via cross-tabulations is distinguished from a more sophisticated one via parametric models.

Basic data analysis strategy

From the perspective of probability theory, each (main) discipline then establishes a random variable. Since certain disciplines are often mentioned together in project proposals (e.g., chemistry and biology), the random variables are usually not independent of each other (similarity), which is expressed in the diversity indicator “disparity” (1-similarity). For the estimation of diversity as entropy, the so-called marginal frequency or share of a discipline is, therefore, not sufficient; cell frequencies are also required (as shown in cross-tabulation of categorical random variables). In the simplest case of 2 disciplines (named or not named in the project application), a 2×2 -cross-tabulation (Fig. 2, above) with cell frequencies results, which can be coded by a simple data design as base for estimation of the corresponding probabilities. Eventually, with cross-tabulations, cell frequencies can already be sufficiently calculated (alternatively via log-linear models).

Sophisticated analysis strategy

Here, a sophisticated analysis strategy is preferred, which offers the opportunity, not only to estimate probabilities, standard errors and confidence intervals, but also allows to include covariates to predict and explain probabilities and to derive statistically diversity measures. For example, it can be tested, whether different funding programs differ in the degree of interdisciplinarity of their funded projects. The central idea of this sophisticated strategy is to express the probabilities p_{ij} (Fig. 2.) in parameters (Eq. 15), which are statistically estimated assuming a certain distribution, here the multinomial distribution.

In order to apply this sophisticated data analysis strategy the data must be organized as in Fig. 2 (below): The different cells of a cross table are designated by k , which are transformed into k 0/1 dummy variables. In order to estimate the variety, the data set is

additionally supplemented by a variable comprising the number of disciplines (NUM_DISC), mentioned in the proposal.

For this data design, a multinomial distribution is ideal as a multivariate distribution of categorical variables, which is a multivariate generalization of the binomial distribution, where $k-1$ cells (and dummy variables) are independent of each other. The following simple multinomial regression model applies (Hedeker, 2003). The k -dimensional vector of dummy variables \mathbf{d} for a unit i (here project application) is multinomially distributed with the k -dimensional vector of cell probabilities \mathbf{p} as parameters.

$$\mathbf{d}_i \sim \text{multinom}(\mathbf{p}_i) \tag{12}$$

however, to allow more complex statistical analysis, the following parametric version is chosen:

$$p_k = \frac{e^{\alpha_k}}{\sum_{k=1}^K e^{\alpha_k}}, \text{ where for } k = K: \alpha_K = 0 \text{ (} e^0 = 1 \text{)} \tag{14}$$

This parameterization offers the opportunity to insert further variables into the model to explain the probabilities, for example a dichotomous variable x_i indicating the funding programme (1 = "interdisciplinary programme", 0 = "other programmes").

$$p_k = \frac{e^{\beta_{0k} + \beta_{1k}x_i}}{\sum_{k=1}^K e^{\beta_{0k} + \beta_{1k}x_i}}, \text{ where for } k = K: \alpha_K = 0 \text{ (} e^0 = 1 \text{)} \tag{15}$$

Data and methods

Data example 1: Metric tide

To answer the question of how diverse research outputs are across research areas, we drew on data from the Metric Tide Report. Metric Tide analyzed the capabilities and limitations of research metrics and indicators. "It has explored the use of metrics across different disciplines and assessed their potential contribution to the development of research excellence and impact. It has analyzed their role in processes of research assessment, including the next cycle of the Research Excellence Framework (REF).

For the analysis, data on submissions for research projects funded since 2006 (Research Excellence Framework 2014) was used. Figure 3 shows an extract of the original cross table comparable to the cross table shown in Table 1 indicating how frequently 20 output types occur in 36 units of assessments (scientific disciplines). It would have been desirable for the analysis to have data on the output types for each individual research project. Such raw data were not available. Furthermore, it is not clear whether the frequencies represent the number of research projects in a cell or the sum of research outputs across all research projects in that cell.

To arrive at relative frequencies or probabilities, according to the simple data analysis strategy the absolute frequencies were divided by the total number ($N = 190.962$). Random variable X is the "output type", and random variable Y is the "research area" (unit of assessment, UOA).

Unit of Assessment	Discipline	Authored book	Edited book	Chapter in Book	Journal article	Conference contribution	Patent / published patent publication	TOTAL
1	Clinical Medicine	0	0	0	13382	7	10	13400
2	Public Health, Health Services and Primary Care	5	0	0	4861	4	0	4881
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	12	6	25	10249	14	15	10358
4	Psychology, Psychiatry and Neuroscience	10	1	16	9086	4	0	9126
5	Biological Sciences	11	0	6	8582	4	3	8608
6	Agriculture, Veterinary and Food Science	1	0	8	3884	7	0	3919
7	Earth Systems and Environmental Sciences	14	4	22	5200	4	3	5249
8	Chemistry	0	0	1	4688	2	3	4698
9	Physics	1	2	1	6376	18	6	6446
10	Mathematical Sciences	46	0	36	6731	17	0	6994
11	Computer Science and Informatics	32	3	112	5551	1898	12	7651
12	Aeronautical, Mechanical, Chemical and Manufacturing Engineering	2	0	9	4101	24	2	4143
13	Electrical and Electronic Engineering, Metallurgy and Materials	0	0	3	3982	28	10	4025
14	Civil and Construction Engineering	3	0	9	1348	16	0	1384
15	General Engineering	7	0	17	8539	90	18	8679
16	Architecture, Built Environment and Planning	229	38	266	2934	77	2	3781
17	Geography, Environmental Studies and Archaeology	380	121	459	4969	23	0	6017
18	Economics and Econometrics	12	0	28	2388	2	0	2600
19	Business and Management Studies	160	6	179	11668	52	0	12202
20	Law	745	25	1219	3454	1	0	5522
21	Politics and International Studies	775	63	415	3082	1	0	4365
22	Social Work and Social Policy	440	34	435	3703	5	0	4784
23	Sociology	350	36	230	2002	1	0	2630

Fig. 3 Extract of the table of frequencies for 23 research areas (UOA)×6 output types (Wilsdon et al., 2015, p. 154)

Data example 2: Ex-ante evaluation data of a Swiss research funding organization

The statistical estimation of diversity and its components is based on data from a Swiss research funding organization. Open access data were downloaded from 11,707 projects (funding type “project funding”) that were funded in the years 2010–2020 (year of project start) (<https://p3.snf.ch/Pages/DataAndDocumentation.aspx>). The analyses were done for the three main disciplines based on 159 subfields: “Biology and Medicine”, “Humanities and Social Sciences”, “Mathematics, Natural- and Engineering Sciences”.

The funding organization does not actually want to make a statement about the interdisciplinarity of its funded projects, but nevertheless codes in the dataset the different disciplines that were relevant in a funded project.

“Researchers allocate their application to a **discipline** in the SNFF list of disciplines. The disciplines are subsumed under research areas, which in turn form three major research domains: humanities and social sciences; mathematics, natural and engineering sciences; biology and medicine. Although many projects are interdisciplinary in nature, we use only the main discipline (as indicated in the application) for the key figures. If an application cannot be allocated to a main discipline based on the available data, we allocate it in equal parts to each of the mentioned disciplines”. (<https://data.snf.ch/key-figures/documentat ion>).

Nevertheless, it is worth analyzing these data for research purposes in order to illustrate the approach described. However, we will refrain from drawing any further conclusions with regard to the funding organization.

The open data policy of the funding organization allows: “Downloading, printing or storage of open data from the SNSF is allowed. If any of this data is published, the source must be cited explicitly.” (<https://p3.snf.ch/Pages/OpenDataPolicy.aspx>). It is also allowed to processing of open data from this funding organization.

Results

Data example 1: metric tide

Different measures can be calculated for the data organized as cross table (Table 5) in order to test, whether research areas differ in their output type or whether output types depend on research areas. Thus, a statistically significant χ^2 -test value, a Cramer’s V of 0.24 and a mutual information of -0.40 show that the output type (X) depends on the research area (Y). Cramer’s V is a correlation measure for categorical variables varying between 0 and 1 and is more or less an alternative measure for the mutual information. There are, as expected, significant differences of the research area in the output types of the research output beyond chance. Looking at the ratio of the unconditional entropy of "output type", $H(X)$, and “research area”, $H(Y)$, to the total entropy, $H(X, Y)$, the differences in frequencies are very much determined by the differences in "research areas" than the differences between "output types". While considering the "research area" reduces about $(1.20-0.80)/1.20=33\%$ of the uncertainty in the "output types", conversely only about $(4.96-4.56)/4.96=8\%$ of uncertainty in the "research areas" is reduced.

Overall, however, the diversity at 10.08 is very high, which is about 72.9% of the maximum possible diversity of 13.81. The maximum variety in the whole table is fully reached and the overall balance of 6.16 is 64.9% of the maximum possible balance of 9.49.

Table 5 χ^2 and entropy measures

Measure	Label	Value	Maximum
χ^2	χ^2 -test value	204,768.3*	
Cramer’s V	Cramer’s V (correlation)	0.24	1.00
$I(X, Y)$	Mutual information (“disparity”)	-0.40	0.00
$H(X)$	Uncond. entropy “output type” (“balance”)	1.20	4.32
$H(X Y)$	Condit. entropy “output type”	0.80	
$H(Y)$	Uncond. entropy “research area” (“balance”)	4.96	5.17
$H(Y X)$	Condit. entropy “research area”	4.56	
$H(X, Y)$	Total entropy	5.76	9.49
<i>Overall indices</i>			
$H_{variety}$	Variety	4.32	4.32
$H_{balance}$	Balance	6.16	9.49
$H_{disparity}$	Disparity	-0.40	0
div_e	Diversity	10.08	13.81

* $p < .05$ (df = 665)

Table 6 Diversity measures for a set of 11 research areas (UOA)

UOA	Name	<i>N</i>	Variety	Balance	Disparity	Diversity
1	Clinical Medicine	4	2.00	0.27	0.02	2.25
2	Public Health, Health Services and Primary Care	4	2.00	0.14	0.01	2.13
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	8	3.00	0.23	0.01	3.22
4	Psychology, Psychiatry and Neuroscience	9	3.17	0.21	0.01	3.37
5	Biological Sciences	7	2.81	0.20	0.01	3.00
6	Agriculture, Veterinary and Food Science	6	2.58	0.12	0.01	2.70
7	Earth Systems and Environmental Sciences	7	2.81	0.15	0.01	2.95
8	Chemistry	5	2.32	0.13	0.01	2.45
9	Physics	8	3.00	0.17	0.01	3.16
10	Mathematical Sciences	9	3.17	0.19	0.01	3.35
11	Computer Science and Informatics	16	4.00	0.23	0.03	4.19

N= number of different output types $\sum(p_i > 0)$

Table 7 Output types with highest individual balance values

Output type	Balance
Journal article	4.15
Chapter in book	0.59
Authored book	0.45
Conference contribution	0.12
Edited book	0.11
Exhibition	0.05

Finally, diversity measures can be derived for the "research areas", which provide information about the diversity for individual research areas. Diversity indicators for the first 11 research areas are shown in Table 6. "Computer Science and Informatics" has the highest diversity with 4.19 and "Public Health, Health Services and Primary Care" the lowest with 2.13.

"Computer Science and Informatics" has the highest variety of 4.0. "Clinical Medicine" has the highest balance (0.27) in this set of selected research areas. Among the different output types, "Journal articles" clearly shows the highest individual entropy, $H(Y)$, i.e., the highest balance among all other output types across different research area (Table 7).

Data example 2: Ex-ante evaluation data

The degree of interdisciplinarity is analyzed at the level of the 3 main disciplines. Overall, almost 90% of the funded projects are disciplinary (Table 8). For the funding instrument "interdisciplinary projects" the percentage of interdisciplinary projects increases to 68.8% compared to 8.7% for the "project funding in the narrow sense". This result provides empirical evidence for the use of these data for measurement issues in the realm of interdisciplinarity. In the following analysis only projects with more than one discipline are selected ($N = 1221$ projects).

Table 8 Cross-tabulation of funding instruments with interdisciplinarity with respect to the 3 main disciplines (absolute frequencies, row percent)

Funding instrument	Interdisciplinary	Disciplinary	Total
“Project funding” in the narrow sense	943 8.7%	9,929 91.3%	10,872 91.9%
“Interdisciplinary projects”	186 68.8%	84 31.1%	270 2.3%
Others	92 16.3%	473 83.7%	565 4.8%
Total	1,221 10.4%	10,486 89.6%	11,707 100%

The statistical approach allows the testing of hypotheses formulated as statistical models. The following hypothesis will be tested (Table 9): There are differences in diversity as measure of interdisciplinarity between projects funded with the funding instrument “interdisciplinary projects” and projects funded with other instruments (M_1). The lower the DIC the better the model fits the data. M_1 shows a lower DIC than the basic model M_0 . There are, actually differences in diversity between the funding instruments “interdisciplinary projects” and the other funding instruments (M_1).

For the basic model (M_0) the different diversity components were estimated with the 95%-credible interval. Additionally, the basic model was estimated for random data, which provide for maximum and reference values of the four different components. The variety of 1 reflect that on the average 2 disciplines were indicated by the applicants ($H_V = -1/\log_2(1/2) = 1$). The balance is about 81% of the maximum possible value, very high. There is some similarity or dependency among the three disciplines, the disparity is not Zero. Diversity is about half of the maximum possible entropy (Table 10).

Table 9 Model comparison regarding the 3 research domains ($N = 1,221$ projects) with the deviance information criterion (DIC)

No	Description	DIC
M_0	Basic model	5,799.2
M_1	“Interdisciplinary projects” vs. other funding instruments	5,792.9

The lower the DIC, the better the model fits the data (best models in italics)

Table 10 Estimated overall diversity components for interdisciplinary projects regarding the 3 research domains ($N = 1,221$ projects) with 95% credible interval in brackets

Model description	Variety	Balance	Disparity	Diversity
Basic Model (M_0)	1.02 [0.96; 1.07]	2.42 [2.35; 2.49]	-0.90 [-0.94; -0.85]	2.55 [2.46; 2.63]
Maximum (random data)	1.59 [1.54; 1.63]	3.00 [2.99; 3.00]	-0.00 [-0.00; -0.00]	4.58 [4.54; 4.62]
Percentages of Maximum	64.2%	80.7%	-	55.5%

Discussion

Diversity is a ubiquitous term used in many disciplines (e.g., ecology, sociology, bibliometrics). A certain break in the discussion on diversity and diversity indicators was brought about by the indicator developed by Rao (1982) and Stirling (2007), which in view of its comparatively simple definition has a very wide circulation. It is to the credit of Rousseau (2019) to point out inconsistencies in this indicator, to which Leydesdorff (2018) has proposed a workable solution. Furthermore, statistical concepts such as probability and correlation/similarity are combined with each other in a way for which there is no statistical basis whatsoever, which was even noted by Rao (1982, p. 7).

Due to the problems of the Rao and Stirling indicator, the aim of this paper was to go back to Shannon's probabilistic concept of entropy (Shannon, 1948), which implicitly dealt with all three facets of diversity, in order to develop a modified concept of diversity from it, which is additive in its nature and allows for both the calculation of a diversity indicator div_e as well as estimation diversity by a statistical model. Due to the fact that most interdisciplinarity indicators trace back to Shannon's entropy, the Shannon's concept of "mutual information" should be favored towards the concept of correlation (Pregowska et al., 2015). A statistical approach was presented how to estimate diversity and its components. In the sense of Occam's razor principle, the question arises as to why this more complex approach should be given preference in practice over the simple concept of Rao and Stirling. First of all, there is no intention to replace the existing diversity indicators, but to identify opportunities for improvement. The following four reasons could be put forward:

1. *Inconsistencies*: Concepts with obvious inconsistencies are not very convincing and reinforce the negative image of bibliometrics.
2. *Interpretation*: Diversity can be interpreted in terms of entropy as a measure of information. Diversity is at its maximum when events can no longer be predicted, as in a coin toss. The lower the diversity, the more predictable events are, the more structure is in the data.
3. *Open the Pandora's box*: The discussions on bibliometric concepts such as field normalization, definition of fields, fractional counting and also diversity seem to be more or less closed with some more or less workable solutions. These considerations might reopen the discussion.
4. *Stochastic nature*: If the occurrence of publications, research output, citations, etc. is assumed to base on a stochastic random process, this must be taken into account within the development of an indicator.
5. *Statistical estimation*: In principle, diversity and its various components can be statistically estimated, as shown.

The calculation of the disparity indicator requires a workable solution on how to deal with the number of components that increases when the number of units increases (e.g., research output types, number of disciplines). For example, with 3 disciplines (A, B, C), there are 4 components of disparity (AB, AC, BC, ABC), with 4 and more units the number of components strongly increases. Eventually, disparity be it mutual information or correlation reflects combinations of disciplines, which often occurs. To get rid of disparity (~ 0) different types of combinations have to distinguished with the help of a latent class analysis (Mutz & Daniel, 2013; Mutz et al., 2012, 2014). Statistically, within a latent class the different units (e.g., disciplines) are uncorrelated.

Acknowledgements The paper is a substantially extended version of a paper, presented at the ISSI 2021 (Mutz, 2021). Figure 2 is an extract of a table from Wilsdon et al. (2015), which is in part copyright HEFCE under Open Government Licence v2.0.

Funding Open access funding provided by University of Zurich.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6), 550–560. <https://doi.org/10.1002/asi.10242>
- Amann, A., & Müller-Herold, U. (2011). *Offene Quantensysteme [Open Quants systems]*. Springer.
- Bornmann, L., Mutz, R., Haunschild, R., de Moya-Anegón, F., Clemente, M. D. M., & Stefaner, M. (2021). Mapping the impact of papers on various status groups in excellencemapping.net: A new release of the excellence mapping tool based on citation and reader scores. *Scientometrics*, 126(11), 9305–9331. <https://doi.org/10.1007/s11192-021-04141-4>
- Brusco, M. J., Cradit, J. D., & Steinley, D. (2020). Combining diversity and dispersion criteria for anti-clustering: A bicriterion approach. *British Journal of Mathematical and Statistical Psychology*, 73(3), 375–396. <https://doi.org/10.1111/bmsp.12186>
- Eshima, N. (2020). *Statistical data analysis and entropy*. Singapore: Springer Nature.
- Foredean, R., Klein, J. T., & Pacheco, R. C. S. (2017). *The Oxford Handbook of Interdisciplinarity*. Oxford University Press.
- Goyanes, M., Demeter, M., Grané, A., Albarrán-Lozano, I., & Gil de Zúñiga, H. (2020). A mathematical approach to assess research diversity: Operationalization and applicability in communication sciences, political science, and beyond. *Scientometrics*, 125(3), 2299–2322. <https://doi.org/10.1007/s11192-020-03680-6>
- Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22(9), 1433–1446. <https://doi.org/10.1002/sim.1522>
- Jones, J. M., & Dovidio, J. F. (2018). Change, challenge, and prospects for a diversity paradigm in social psychology. *Social Issues and Policy Review*, 12(1), 7–56. <https://doi.org/10.1111/sipr.12039>
- Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10), 2427–2439. <https://doi.org/10.1890/06-1736.1>
- Kruschke, J. K. (2011). *Doing Bayesian data analysis - A tutorial with R and BUGS*. Elsevier.
- Leydesdorff, L. (1995). *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communication*. Leiden: DSWO, Leiden University.
- Leydesdorff, L. (1991). The static and dynamic analysis of network data using information theory. *Social Network*, 13, 301–345.
- Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics*, 116(3), 2113–2121. <https://doi.org/10.1007/s11192-018-2810-y>
- Leydesdorff, L., & Ivanova, I. (2021). The measurement of “interdisciplinarity” and “synergy” in scientific and extra-scientific collaborations. *Journal of the Association for Information Science and Technology*, 72(4), 387–402. <https://doi.org/10.1002/asi.24416>
- Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019a). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics*, 13(3), 904–905. <https://doi.org/10.1016/j.joi.2019.03.016>
- Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019b). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1), 255–269. <https://doi.org/10.1016/j.joi.2018.12.006>

- Mutz, R. (2021). *Diversity and interdisciplinarity - Should variety, balance and disparity be combined as a product or better as a sum? A probability-theoretical approach*. Paper presented at the 18th International Conference on Scientometrics and Informetrics, ISSI 2021.
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Types of research output profiles: A multilevel latent class analysis of the Austrian Science Fund's final project report data. *Research Evaluation*, 22(2), 118–133. <https://doi.org/10.1093/reseval/rvs038>
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2014). Cross-disciplinary research: What configurations of fields of science are found in grant proposals today? *Research Evaluation*, 24(1), 30–36. <https://doi.org/10.1093/reseval/rvu023>
- Mutz, R., & Daniel, H. D. (2013). University and student segmentation: Multilevel latent-class analysis of students' attitudes towards research methods and statistics. *British Journal of Educational Psychology*, 83(2), 280–304. <https://doi.org/10.1111/j.2044-8279.2011.02062.x>
- Pregowska, A., Szczepanski, J., & Wajnryb, E. (2015). Mutual information against correlations in binary communication channels. *BMC Neuroscience*. <https://doi.org/10.1186/s12868-015-0168-0>
- Rao, R. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: the Indian Journal of Statistics, Series A (1961-2002)*, 44(1), 1–22.
- Rousseau, R. (2018). The repeat rate: From Hirschman to Stirling. *Scientometrics*, 116(1), 645–653. <https://doi.org/10.1007/s11192-018-2724-8>
- Rousseau, R. (2019). On the Leydesdorff-Wagner-Bornmann proposal for diversity measurement. *Journal of Informetrics*, 13(3), 906–907. <https://doi.org/10.1016/j.joi.2019.03.015>
- Rousseau, R., Van Hecke, P., Nijssen, D., & Bogaert, J. (1999). The relationship between diversity profiles, evenness and species richness based on partial ordering. *Environmental and Ecological Statistics*, 6(2), 211–223. <https://doi.org/10.1023/A:1009626406418>
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shaw, W. M., Jr. (1981). Information theory and scientific communication. *Scientometrics*, 3(3), 235–249. <https://doi.org/10.1007/BF02101668>
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719. <https://doi.org/10.1098/rsif.2007.0213>
- Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0127298>
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Johnson, B. (2015). *The Metric Tide: Report on the independent review of the role of metrics in research assessment and management*. Retrieved from https://responsiblemetrics.org/wp-content/uploads/2019/02/2015_metrictide.pdf