A scientometric analysis of diversity in HEP over the past three decades

Maria Grazia Pia (INFN Genova, Italy)  
maria.grazia.pia@cern.ch

Tullio Basaglia (CERN), Zane W. Bell (ORNL),  
Arnold Burger (Fisk. Univ.), Paul V. Dressendorfer (IEEE)

ICHEP 2018  
Seoul, 4-11 July 2018
Diversity from a scientometric perspective

Scientometrics
The study of measuring and analysing science

Scope of this study
Publications in scholarly journals, 1985-2017
Particle physics, Nuclear physics, Astrophysics, HEP technology

Data sources
Web of Science (Clarivate Analytics)
UN and OECD data

Countries, organizations, authors, journals

Observables
Descriptive statistics
Measures of diversity
Measures of inequality

Statistical methods
Statistical inference: trend tests, correlation tests
Caveat

- CERN cancelled the subscription to the Web of Science (1970→) in July 2017
- It is no longer possible to do any scientometric studies at CERN
- WoS access through INFN (1990→) and Univ. of Genova (1985→)

The WoS is affected by several known problems too many to mention, specific warnings in the course of the presentation

WoS categories

- Physics, **Particles & Fields** (excluding NIM A)
- Physics, **Nuclear**  
  - Astronomy & Astrophysics  
    - no conference papers
- **HEP technology** journals: NIM A/B, IEEE TNS, JINST

Maria Grazia Pia, INFN Genova
33 years of HEP publications

General increase of the number of publications, of journals, of participating countries, organizations and authors
Compared to other research domains

- **Publications**: Astrophyics, Nuclear Physics, Particle Physics, HEP Technology
- **Countries**: Astrophyics, Nuclear Physics, Particle Physics, HEP Technology
- **Journals**: Astrophyics, Nuclear Physics, Particle Physics, HEP Technology
- **Organizations**: Astrophyics, Nuclear Physics, Particle Physics, HEP Technology
- **Authors**: Astrophyics, Nuclear Physics, Particle Physics, HEP Technology

**Beware**: nuclear physics and astrophysics data sets exclude journals also classified as particle physics.
Data distributions

Contradiction: low and approximately constant median, outliers extending up to very large number of publications

Food for thought: scientific and sociological implications
The world changes…

Particle physics publications

Same set of countries as the WoS particle physics data
Particle physics

Top 10

Particle/nuclear technology
Rank in particle physics

USA

Germany

Switzerland

Russia

Brazil

Argentina

Colombia

Mexico

South Africa

Tunisia

Iran

Israel
$H_0$: no trend, $H_1$: p < 0.01
Diversity

Ecological diversity measures richness and complexity of a community.

Species diversity measures the different kinds of organisms living in a community.

- Related to the concept of entropy in information theory.
- Measured by several indices, with different sensitivity to rare species.
  - Mathematical functions that combine richness and evenness in a single measure.
  - Correlations between indices due to their common background.
  - Different measures may be appropriate, depending on the context.
Diversity measures

**Margalef index**

$$D_{Mg} = \frac{(S - 1)}{\log N}$$

**Simpson index**

$$\lambda = \sum_{i=1}^{R} p_i^2$$

**Gini-Simpson**

$$D_1 = 1 - \sum_{i=1}^{S} p_i^2$$

**Shannon index**

$$H = -\sum_{i=1}^{S} p_i \log p_i$$

**Renyi diversity**

$$q H_{Renyi} = \frac{1}{1-q} \log \left( \sum_{i=1}^{S} p_i^q \right)$$

**Tsallis diversity**

$$q H_{Tsallis} = \frac{1}{1-q} \left( \sum_{i=1}^{S} p_i^q - 1 \right)$$

**Hill numbers**

$$q D = \left( \sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}$$

- **0**: species richness
- **1**: exp(Shannon)
- **2**: $$\lim_{q \to 1} q\ D = 1/\sum_{i=1}^{S} p_i^2$$

**Diversity indices**

**Rarefaction**

**Taxonomic and functional diversity**

- **Average distance of traits**
- **Height of trait tree**

**Species abundance models**

- **Fisher and Preston**
- ** Ranked abundance distribution**

**Species accumulation and beta diversity**

- **Species accumulation models**
- **Beta diversity**

**Species pool**

- **Number of unseen species**
- **Pool size from a single site**
- **Probability of pool membership**

**The vegan package** has two major components: multivariate analysis (mainly ordination), and methods for diversity analysis of ecological communities. This document gives an introduction to the latter. Ordination methods are covered in other documents. Many of the diversity functions were written by Roeland Kindt, Bob O’Hara and Péter Sólymos.

Most diversity methods assume that data are counts of individuals. The methods are used with other data types, and some people argue that biomass or cover are more adequate than counts of individuals of variable sizes. However, this document mainly uses a data set with counts: stem counts of trees on 1 ha plots in the Barro Colorado Island. The following steps make these data available for the document:

```R
> library(vegan)
> data(BCI)
```

### Abstract

This document explains diversity related methods in vegan. The methods are briefly described, and the equations used them are given often in more detail than in their help pages. The methods discussed include common diversity indices and rarefaction, families of diversity indices, species abundance models, species accumulation models and beta diversity, extrapolated richness and probability of being a member of the species pool. The document is still incomplete and does not cover all diversity methods in vegan.
Species richness
\((\text{Hill}_0)\)

Margalef

Shannon

HEP technology

Maria Grazia Pia, INFN Genova
Renyi diversity

Extension of Shannon entropy

The higher the order \((q)\), the lower the sensitivity to rare species

\[
q H_{\text{Renyi}} = \frac{1}{1 - q} \log \left( \sum_{i=1}^{S} p_i^q \right)
\]

order = 1 (Shannon)

order = 2

order = \(\infty\)
Renyi diversity

Similarity across all disciplines

Lower values at higher order, where rare species (organizations) have lower impact

Trend: slowly increasing
Renyi diversity

Lower values at higher order, where rare species (*journals*) have lower impact

Trend

order = 0

order = 1

order = 2

order = ∞
General consensus reached relatively recently on Hill numbers as measures of diversity.

Hill number of order 1

$\exp(\text{Shannon})$
All that glitters…

Large and generally increasing diversity in HEP publications, but…

How fairly are scholarly publications distributed within the scientific community?

How does their distribution evolve as a function of time?

Econometric analysis

Inequality

Trend
Most common measure of inequality

\[ G = 2 \int_0^1 [x - L(x)] \, dx \]

“\( x \) richest people in the world are worth more than the poorest \( y \)%”

0 \( \leq G \leq 1 \)

more unequal society

C. Gini, Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche, 1912
**Pietra index**

\[ P = \max (L_{pe}(x) - L(x)) \]

- Used in derivative markets as a benchmark measure of statistical heterogeneity
- Counterpart of Kolmogorov-Smirnov statistic
- It can be interpreted as the proportion of income that has to be transferred from those above the mean to those below the mean in order to achieve an equal distribution

*AKA Ricci-Schutz index, Hoover index, Robin Hood index*
Other inequality measures

Theil index

\[ T = \sum_{i=1}^{n} s_i \left[ \log s_i - \log \left( \frac{1}{n} \right) \right] \]

- \( s_i \) = share of the \( i \)th group in total income
- \( n \) = total number of income groups

More equal society

The same as redundancy in information theory: the maximum possible entropy of the data minus the observed entropy

Atkinson index

\[ I = 1 - \frac{\pi_e}{\mu} \]

- \( e \) = sensitivity parameter
- \( 0 \leq I \leq 1 \)

Used to calculate the proportion of total income that would be required to achieve an equal level of social welfare as at present, if incomes were perfectly distributed...
Inequality evolution

Trend tests
Mann-Kendall, Cox-Stuart \( \alpha = 0.01 \)

Inequality decreases across countries, but HEP publications are more and more concentrated within a small number of organizations.

Similar trends in Gini, Pietra and Atkinson indices

\( H_0: \) no trend
\( H_1: \) trend

\( p < 0.001 \)
Inequality among HEP authors has been increasing over time.

No univocal indication of general trend regarding HEP journals.
Organization and author inequality in HEP technological publications is lower than in physics publications

Particle physics exhibits the largest inequality among authors, HEP technology the lowest

\[ p < 0.01 \]

<table>
<thead>
<tr>
<th>( H_0 ): no trend</th>
<th>( H_1: \uparrow ) trend</th>
<th>( H_1: \downarrow ) trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Particle</td>
<td>organization</td>
<td>author</td>
</tr>
<tr>
<td>HEP technology</td>
<td>author</td>
<td>country</td>
</tr>
<tr>
<td>Astro</td>
<td>country</td>
<td>organization</td>
</tr>
<tr>
<td>Nuclear</td>
<td>organization</td>
<td>author</td>
</tr>
</tbody>
</table>
Conclusions

Publication patterns in particle physics and related disciplines are studied with statistical analysis methods derived from ecology and econometrics.

In general, evolution towards greater diversity in HEP, but increasing concentration in a small number of organizations.

Food for thought: evolution of authorship and its meaning.

“Biodiversity is the totality of all inherited variation in the life forms of Earth, of which we are one species. We study and save it to our great benefit. We ignore and degrade it to our great peril.” E.O. Wilson