*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Manuscript Category

|  |
| --- |
| Bioinformatic Section  Pre-processing of data in biodiversity analysis: a Python beginner simulation.  Jesus David Sierra1,\*  1Department of Genetics and Evolution, Federal University of São Carlos, Rod. Washington Luís 235. São Carlos, Brazil.  \*jesussierra@estudante.ufscar.br  Abstract  **Motivation:** Data pre-processing is a critical stage in the analysis of biological diversity of any community; the veracity of the estimates and the understanding of how that diversity is formed depend on the quality of the data that will be analyzed. In this work, three forms of species composition data are analyzed quickly and simply to note the differences caused by pre-processing in the estimation of biological diversity.  **Results:** Pre-processing treatments show notable differences from the original dataset. Rarefaction makes each sample the same size (42242 reads), while normalization adjusts data to a consistent scale (86720-86730 reads), preserving variations. Rarefaction facilitates sample comparison, but normalization retains rare species information. Treatments slightly affect the result of Shannon index. Both treatments are essential techniques for making meaningful comparisons between datasets, ensuring accurate and unbiased biodiversity assessments.  **Availability:** https://github.com/Jesus-sierra/Bioinformatic\_course\_simulation  **Contact:** jesussierra@estudante.ufscar.br  **Supplementary information:** Supplementary data are available at previous GitHub informed. |

# Introduction

Biological data, particularly species composition data, are essential for understanding biodiversity within ecosystems. These data are typically presented in tables that list species observed in different samples or locations, along with their respective abundances or occurrences. Accurate analysis of these data requires methods to account for sampling effort and ensure comparability across datasets. Two key techniques used in this context are rarefaction and normalization1.

Rarefaction is a statistical method used to equalize the species richness across samples that have different sampling efforts, so that the species diversity can be compared by calculating an expected number of species in smaller standardized sample size. This is particularly beneficial when some samples have been more intensively sampled than others. The process of rarefaction implies randomly drawing subsets from the dataset for multiple times, placing the mean number of observed species vs number of individuals sampled, then, a rarefaction curve is made to estimate species richness for a standardized sample size1,2.

Normalizations adjust species abundance data for differences in sampling weights or any other factors that may influence the count of species. It is significant to note that this approach ensures data on comparable scales, and this is important when carrying out biodiversity analysis as well as rarefaction. However, normalization could be achieved by different methods. The Relative Abundance Transformation convert raw abundance data into relative proportions by dividing each species count by the total count in that sample. Besides, Log Transformation apply a logarithmic transformation to abundance data to reduce the influence of highly abundant species and highlight the presence of less common species. Finally, Standardization adjust species counts based on sample area or volume, making the data comparable across samples of different sizes. These methods further increase the variability in the results that a biological data set may have when analyzed using normalization3.

Through employing rarefaction/normalization, ecologists can compare among sites or samples for species richness without being biased by varying levels of intensities within those sample areas. Tables of species composition serve as a fundamental tool for conducting ecological studies. They list the presence and abundance of species across various sites or samples, facilitating the assessment of biodiversity patterns2,4.

The aim of this work was to generate tables of species composition with different pre-process (raw, rarefied and normalized) to be analyzed in order to identify differences caused by pre-processing in the estimation of biological diversity while modern computing tools such as Python are put into practice.

# Methods

## Table generation

A table of OTUs was generated in TSV format, consisting of 26 samples (columns) and an index with 100 OTUs ('OTU n'). Considered as original table. Each sample in this table contains at least 40% and at most 75% of cells as zeros and adds up to a maximum of 100,000 counts, simulating reads obtained from a sequencing process. Script available in supplementary material as makeotutable.py.

Furthermore, a rarefied table was generated from the original table of OTUs following the principle of random selection of values, divided by the minimum number of counts for each sample. Script available in supplementary material as rarefy.py.

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamenteAlternatively, a normalized table was created converting each OTU to its proportion in the sample and then multiplying it by a large number (maximum reads value of each sample). Script available in supplementary material as TSS.py.

Finally, a table was created containing the mean values ​​of counts per sample and standard deviation of the means for each of the treatments (original, rarefied and normalized). Script available in supplementary material as means\_table.py.

## Visual pre-process comparison

Boxplot graph was created to show the sum of reads per sample using Seaborn and Matplotlib resources, therefore visual characterization between treatments could be made. Script available in supplementary material as sum\_table\_plot.py

## Collector’s curve analysis

Collector's curve is an important analysis that could determine the efficiency of sampling, for biological diversity it is vital. Unfortunately, Collector's curve analysis was not achieved in the present work.

## Shannon index analysis

Shannon index comparison was made to determine influence of preprocessing treatments in diversity representation. Thus, a script containing de Shannon function was created based on a simply process a) OTUs values in each sample were divided by the total number of counts to obtain its proportion in the sample, b) OTUs proportion were multiplied by their own logarithm, c) resulting values ​​were added and then multiplied by -1, thus obtaining a positive value for each sample (Shannon index). Shannon index script is available in supplementary material as shannon.py.

# Results

The generation of tables with Python and its Pandas and NumPy packages is very effective when carrying out simulations that allow decisions to be made when analyzing biological data sets. In the same way, when having an original dataset, rarefaction and/or normalization are simple and quick processes to perform with Python.

The visual comparison of the pre-processing treatments shows a great difference compared to the original dataset (Figure 1). The rarefaction of the table offers a strong fit to the data, where the result of the sum of reads is the same for each sample (42242) leaving all samples the same size. This is particularly good because it creates stability in relation to sampling effort or sequencing depth.

In the case of normalization, it is possible to observe that the data set was adjusted to the same scale with reads count values ​​ranging between 86720 and 86730 for all samples. This discrete variation in the data shows a great difference between the two methods tested, this means, while rarefaction equalizes the size of the samples to allow comparison between them, normalization preserves the variations between them, understanding that each sample has its peculiarities. Thus, the normalization method can avoid the loss of information when it comes to rare species, contrary to rarefaction in which these elements could be underestimated.

It is worth mentioning that the average results for the untreated (original) data set show a variation that makes direct comparison of the information contained impossible and makes evident the need to opt for one or more pre-processing method. Variations on means per sample are summarized in Table 1.

**Fig. 1. Visual comparison of treatments.** No pre-processed data (original) show by far large variation, rarefied data counts the same number reads per sample and normalized data show minimal variation on data counts per sample.

**Table 1.**Means of sample counts per treatment

|  |  |  |
| --- | --- | --- |
| Treatment | Mean | STD |
| Original | 6604726923076920 | 12702454837119200 |
| Rarefied | 42242.0 | 0.0 |
| Normalized | 8672553846153840 | 38950735080529700 |

Obtained with NumPy (Python3).

Shannon index results are listed and available in supplementary material and are of importance because it allows us to compare how the diversity of a sample varies according to the pre-processing carried out. Figure 2 show treatments slightly affect the result of Shannon index. Unfortunately, carrying out of time difficult to continue analyzing these data.

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente**Fig. 2. Shannon index representation.** Proximity in data show slightly variation between treatments.

Understanding and analyzing biodiversity through biological data and species composition tables is fundamental in ecology. Rarefaction and normalization are essential techniques that allow for meaningful comparisons between datasets, ensuring that biodiversity assessments are accurate and unbiased. By applying these methods, it is possible to gain deeper insights into the structure and function of communities in the most diverse ecosystems.

Acknowledgements

To the Postgraduate Program in Ecology and Natural Resources at the Federal University of São Carlos UFSCar for offering the Python course for bio scientists.

Funding

This work has been supported by the National Council for Scientific and Technological Development (CNPq)

*Conflict of Interest:* none declared.

References

1. Gotelli, N. J., & Colwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4), 379-391.
2. Colwell, R. K., & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London*. Series B: Biological Sciences, 345(1311), 101-118.
3. Lin, H., Peddada, S.D. (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *npj Biofilms Microbiomes* 6, 60. https://doi.org/10.1038/s41522-020-00160-w
4. Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4), 783-791.