Coverage-based Comparisons of Cultural Diversity

Folgert Karsdorp, Melvin Wevers, Mike Kestemont

Background

Collections of cultural heritage material are known to contain biases. These biases can influence our analyses based on these collections. One such bias pertains the diversity in a collection, i.e. how much variation and repetition can we find in a collection. In this paper, we quantify diversity in cultural collections, and reveal how this presents unique challenges, particularly when assessing the diversity across different collections. We focus on the complexities of comparing cultural diversity in a heritage context, using a case study of a unique collection of Dutch folktales assembled by the Meertens Institute in the 1960s and 70s (Meder et al. 2016).

Methods

This study's primary method assesses the 'richness' of the collection, defined as the count of unique items or categories present. This is initially measured by the total number of different folktale types within the collection, as classified by standard folklore taxonomies such as those by (Uther 2004).

To address the limitations of richness as a sole measure, particularly its sensitivity to sample size, the study incorporates the concept of rarefaction, borrowed from ecological research (Chao and Jost 2012). Rarefaction is a technique used to standardize different sample sizes, allowing for a fair comparison of diversity across collections. It involves downsizing larger collections to match the size of smaller ones and computing the richness for these standardized sizes. The process generates rarefaction curves that depict the relationship between the number of unique categories and the sample size. This method provides a more nuanced understanding of diversity by adjusting for varying sampling efforts and collection sizes.

Further, the study introduces a novel approach of standardizing samples based on 'coverage' rather than size. Coverage, as defined in the study, refers to the proportion of the total diversity represented in a sample (Good 1953). This method involves estimating the sample coverage using equations derived from the field of cryptanalysis—a process of finding weaknesses in cryptographic algorithms—and ecological statistics. The use of coverage-based rarefaction curves allows for a comparison of collections at equivalent levels of completeness, providing a more accurate representation of their relative diversity (Alroy 2010; Chao and Jost 2012).

Results

The study demonstrates significant differences in the diversity measurements when standardizing by sample size versus coverage. In the case of Dutch folktales, collections from different collectors like Jaarsma and Kooijman showed varying richness due to differences in sampling effort. The application of rarefaction curves revealed that the apparent diversity differences

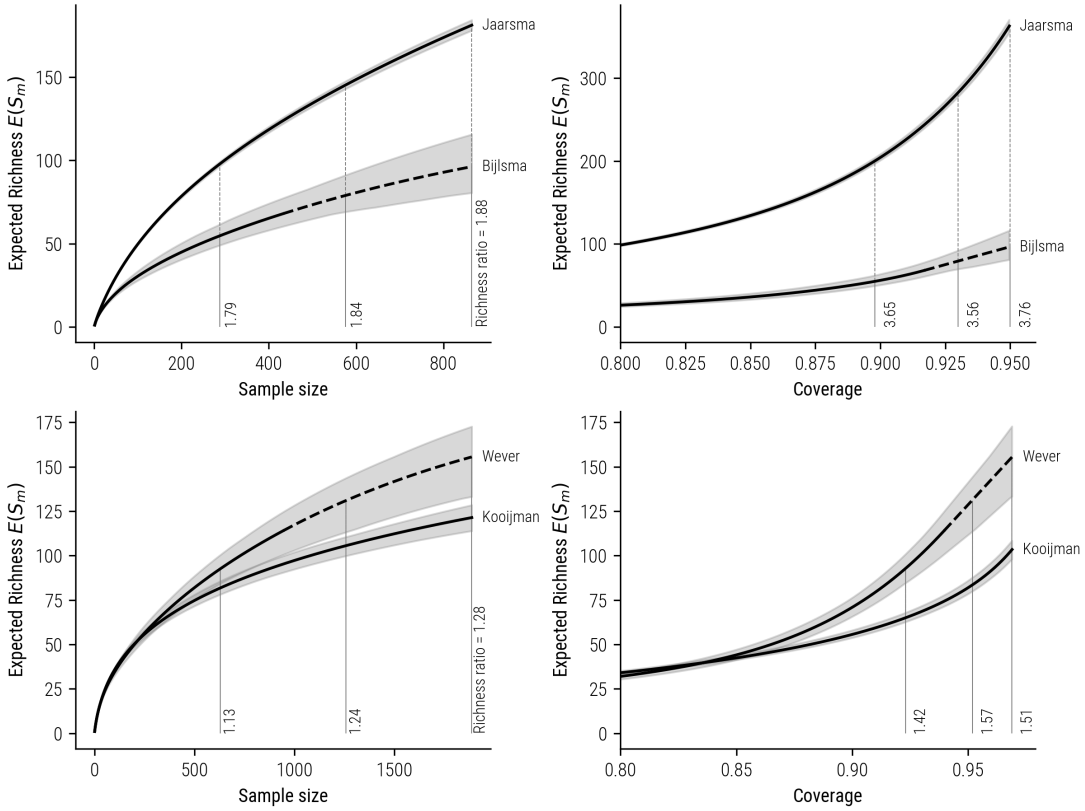

Figure 1: Rarefaction-extrapolation curves based on sample size (left) compared with coverage-based curves for the collectors Jaarsma, Bijlsma, Wever and Kooijman.

were more a function of the chosen sample size. Furthermore, standardizing by sample coverage provided a more accurate reflection of the actual diversity relationship between collections. This method adhered to the doubling property (Hill 1973), suggesting its effectiveness in measuring diversity.

Discussion

The findings underscore the importance of considering sample coverage in diversity estimates. The traditional richness measurements, based solely on sample size, often fail to capture the true diversity due to their sensitivity to sampling effort. By standardizing based on sample coverage, this study demonstrates a more accurate and fair comparison of diversity across collections. This approach not only allows for a more precise measurement of diversity within a single collection but also provides a reliable basis for comparing diversity across different collections.

This research has significant implications for cultural studies, particularly in the field of archival studies and heritage conservation. The methodology developed can be used to assess and compare the cultural diversity of various collections across the GLAM sector more accurately, allowing for better understanding and preservation strategies. It also opens new avenues for further research into the factors influencing coverage and diversity in cultural collections.

In conclusion, this study contributes to the field of cultural diversity measurement by introducing a novel approach that considers sample coverage for more accurate comparisons. The application of this method to Dutch folktale collections not only highlights the nuances in

cultural diversity assessment but also sets a precedent for future studies in similar domains.

References

Alroy, John. 2010. "Geographical, Environmental and Intrinsic Biotic Controls on Phanerozoic Marine Diversification: Controls on Phanerozoic Marine Diversification." Palaeontology 53 (6): 1211–35. https://doi.org/10.1111/j.1475-4983.2010.01011.x.

Chao, Anne, and Lou Jost. 2012. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size." Ecology 93 (12): 2533–47. https://doi.org/10.1890/11-1952.1.

Good, I. J. 1953. "The Population Frequencies of Species and the Estimation of Population Parameters." Biometrika 40 (3-4): 237–64. https://doi.org/10.1093/biomet/40.3-4.237.

Hill, M. O. 1973. "Diversity and Evenness: A Unifying Notation and Its Consequences." Ecology 54 (2): 427–32. https://doi.org/10.2307/1934352.

Meder, Theo, Folgert Karsdorp, Dong Nguyen, Mariët Theune, Dolf Trieschnigg, and Iwe Muiser. 2016. "Automatic Enrichment and Classification of Folktales in the Dutch Folktale Database." Journal of American Folklore 129 (511): 78–96.

Uther, Hans-Jörg. 2004. The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson. 284-286. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.