Forgotten books : the application of unseen species models to the survival of culture

Reference:
Kestemont Mike, Karsdorp Folgert, de Bruijn Elisabeth, Driscoll Matthew, Kapitan Katarzyna A., Ó Macháin Pádraig, Sawyer Daniel, Sleiderink Remco, Chao Anne.- Forgotten books : the application of unseen species models to the survival of culture
Science / American Association for the Advancement of Science [Washington, D.C.]- ISSN 1095-9203 - 375:6582(2022), p. 765-769
Full text (Publisher's DOI): https://doi.org/10.1126/SCIENCE.ABL7655
To cite this reference: https://hdl.handle.net/10067/1859430151162165141

Title: Forgotten Books: The Application of Unseen Species Models to the Survival of Culture

Authors: Mike Kestemont,[1*] † Folgert Karsdorp,[2] † Elisabeth de Bruijn,[1,3] Matthew Driscoll,[4] Katarzyna A. Kapitan,[5−9] Pádraig Ó Macháin,[10] Daniel Sawyer,[11] Remco Sleiderink,[1] Anne Chao[12]

Affiliations:

[1]University of Antwerp; Antwerp, Belgium

[2]KNAW Meertens Institute; Amsterdam, the Netherlands

[3]Ruhr-Universität Bochum; Bochum, Germany

[4]Arnamagnæan Institute, University of Copenhagen; Copenhagen, Denmark

[5]Linacre College, University of Oxford; Oxford, United Kingdom

[6]Department of Nordic Studies and Linguistics, University of Copenhagen; Copenhagen, Denmark

[7]Vigdís Finnbogadóttir Institute of Foreign Languages, University of Iceland; Reykjavík, Iceland

[8]The National Museum of Iceland; Reykjavík, Iceland

[9]The Museum of National History, Frederiksborg Castle; Hillerød, Denmark

[10]University College Cork; Cork, Ireland

[11]Merton College, University of Oxford; Oxford, United Kingdom

[12]Institute of Statistics, National Tsing Hua University; Hsin-Chu, Taiwan

*Corresponding author. E-mail: mike.kestemont@uantwerp.be

†These authors contributed equally to this work.

Abstract: The study of ancient cultures is hindered by the incomplete survival of material artefacts, so that we commonly underestimate the diversity of cultural production in historic societies. To correct this survivorship bias, we apply unseen species models from ecology to gauge the loss of narratives from medieval Europe, such as the romances about King Arthur. The estimates obtained are compatible with the scant historic evidence. Besides events like library fires, we identify the original evenness of cultural populations as an overlooked factor in these assemblages' stability in the face of immaterial loss. We link the elevated evenness in island literatures to analogous accounts of ecological and cultural diversity in insular communities. These analyses call for a wider application of these methods across the heritage sciences.

One-sentence summary: Unseen species models from ecology can estimate artefact survival rates from ancient cultures.

Main Text: Historical studies of human culture are hindered by the fact that they must work with incomplete samples of material artefacts (books, paintings, statues etc.) that still survive (*1*, *2*) but do not necessarily represent the original population faithfully. Because of this survivorship bias, we risk underestimating the diversity of the cultural production of past societies. In response to this risk, we turn to bias correction methods from ecology. For monitoring species richness reliably, ecologists use statistical models that account for the unseen species in samples (*3*). This is necessitated by the common under-detection of species that are hard to observe during bioregistration campaigns, creating a detection bias that must be quantitatively accounted for. Following recent studies (*4*, *5*) pointing to parallels between cultural and ecological diversity, we show that unseen species models can be applied to manuscripts preserving medieval literature. This enables us to estimate the size of the original population of works and documents and, in turn, the losses that these cultural domains sustained. We offer a large-scale estimate of the (im)material loss of narrative fiction from medieval Europe. This endeavor resonates with a broader interest in the persistence of cultural information in human societies, particularly in the domain of cultural evolution (*5–9*).

Fig 1. *Top left (A)*: Fragment of *Strengleikar* (COPENHAGEN, DEN ARNAMAGNÆANSKE SAMLING, AM 666 B 4TO), repurposed to stiffen a bishop's miter. Used with permission. *Top right (B)*: Intact, lavishly illustrated codex (*Wigalois*; LEIDEN, UNIVERSITY LIBRARY, LTK. 537, F. 72V). CC-BY. *Bottom (C)*: Fragment (binding waste) of an unidentified Dutch romance (KU LEUVEN LIBRARIES, SPECIAL COLLECTIONS, MS. 1488). Public domain.

Narrative fiction was a mainstay of medieval culture (ca. 600–1450 CE). The courtly chivalric romance, for example concerning King Arthur and the Knights of the Round Table, has had a long-lasting impact. Before movable-type printing in Europe (ca. 1450 CE), handwritten documents (manuscripts) were used for the sustainable storage of text (*10*). In some places – e.g., Ireland and Iceland – manuscript circulation continued in this role into the modern era. Works of narrative fiction circulated through manually produced copies that survive as unique material artefacts, typically in the form of parchment or, later, paper codices (*11*). Thus, multiple parallel witnesses of the same medieval work could circulate. Today, manuscripts constitute the main evidence regarding medieval narrative fiction. Textual witnesses have been subject to various processes of decay and destruction (e.g., library fires) (*1*, *2*, *11*, *12*). Texts may survive in intact codices (Fig. 1B), but many of those works which survive at all now only exist in manuscripts that are fragmentary, lacking leaves or bearing damage from tearing, insects, overuse etc. Because of parchment's durability, books were often recycled for more everyday practical uses (Fig. 1A), e.g. made into small boxes, used as tailors' measures or even packing material for meat. Additionally, strips of parchment were frequently used by binders to strengthen book spines (Fig. 1C).

The (material) loss of documents can entail the (immaterial) loss of works: a work becomes "lost" when none of the copies that once preserved it are known to have survived (*13*). A theoretical distinction must be made between documents that have been destroyed and those which have not been recovered yet, e.g. because of inadequate cataloguing: sources in the latter category might still reemerge. Different survival scenarios are represented in Fig. 2. We adopt a distinction between the (non-material) WORK, as listed in pre-existing scholarly repertories, and the (material) DOCUMENTS in which these WORKS are attested (*14*). While medieval narratives also circulated orally, the present analysis is necessarily limited to written production.

Fig 2. Schematic representation of example survival scenarios (*A*–*G*) for medieval literature. Individual WORKS were copied into one (*A*–*E*) or more (*F*–*G*) DOCUMENTS, whose survival STATUS varies from intact codices (*A*) to fragments (*C*, *E*), residing in REPOSITORIES, such as libraries, archives or private collections. Lost DOCUMENTS can be fully (*D*) or partly (*G*) destroyed, or may not have been recovered yet (*B*). For lost WORKS (*B*, *D*), none of the original documents have been recovered.

The survival rates for medieval DOCUMENTS are traditionally estimated based on medieval library catalogues: if the listed specimens can still be identified, the calculation of the survival rates of these books is straightforward (*1*). Authoritative studies have suggested (for the Holy Roman Empire) an overall survival rate of ~7% for general-purpose manuscripts, which must be adjusted upwards to ~20% for higher-end codices (*1*, *11*, *15*). Such estimates are nevertheless problematic because they depend on a small sample of catalogues, from protected collection environments, with cataloguers frequently omitting lower-end documents (*15*). A prior attempt (*16*) to apply methods from survival studies to this problem met with criticism, as the figures obtained did not fit with other historical evidence (*17*, *18*). Regarding the loss of WORKS, there has been little quantitative work (*19*). Conventional approaches rely on allusions to lost works, e.g., in library catalogues (*13*), but many lost works will not have been mentioned. Egghe and Proot published a pioneering estimator for the loss of multi-copy, printed works (*20*), which was later identified as an unseen species model. Their approach, however, requires an estimate of the print-runs of hand-pressed books, which does not suit manuscripts.

We build on the information-theoretic analogy that medieval WORKS can be treated as distinct species in ecology, and that the number of extant DOCUMENTS for each WORK can be regarded as analogous to the number of sightings for an individual species in a sample. Thus, if we treat the available count information for medieval literature as "abundance data" (*3*), one can apply unseen species models to estimate the number of lost WORKS in a corpus or assemblage. We collected count data for surviving medieval heroic and chivalric fiction in six European vernaculars (*21*): three insular (Irish, Icelandic, English) and three continental (Dutch, French, German). For all WORKS, we listed the number of handwritten medieval DOCUMENTS in which they survive (Table 1). Next, we applied non-parametric methods to estimate the original richness of these traditions. For a given assemblage, let $(X_1, X_2, \ldots, X_{S_{obs}})$ represent the abundance-based frequencies for S_{obs} unique works which were observed in n documents.

Chao1 is a method to estimate a lower bound on $\hat{f}_0$, or the number of undetected species in an assemblage, based on the number of singletons (f_1, species sighted only once) and doubletons (f_2, species sighted exactly twice) in a sample of n individuals. The original number of works ($\hat{S}$) can then be estimated as $S_{obs} + \hat{f}_0$ (*22*). Chao1 is not specific to ecology and has been derived under a very general model: it can be applied as a universally valid lower-bound richness estimator to any hyper-diverse, under-sampled collection of types, such as stone tools, coins or even words (*23*). Therefore, this estimator is even more widely applicable in the heritage sciences than shown here (*24*). In this framework, the survival ratio for the WORKS can be quantified as the *sample completeness* or $S_{obs}/\hat{S}$: the ratio of the number of unique observed works (S_{obs}) over the estimated true species abundance $\hat{S}$ (*25*). Species richness is an intuitive measure to quantify species diversity, but alternative measures exist, e.g. the Shannon or Simpson diversity (both put less weight on rare species). The Hill number profile (*26*) allows us to compare a sample's diversity across various values of q, a scalar corresponding to different diversity measures at specific points (e.g., $q = 0$ for richness, $q = 1$ for Shannon, $q = 2$ for Simpson). Hill numbers are nowadays the diversity measure of choice in ecology for quantifying species diversity and decomposition (*25*).

We also use an extension of Chao1 (*27*) that estimates the minimum number m of additional observations that are required to observe each of the $\hat{f}_0$ species at least once. This number will approximate the number of lost DOCUMENTS in an assemblage, so that we can estimate the original population size as $n + m$. Chao1 and the minimum sampling extension were derived as a *lower* bound, which implies that the estimates of the survival ratios below, strictly speaking, offer an *upper* bound on the loss of WORKS and DOCUMENTS – it is possible that even more literature was lost. Nevertheless, Chao1 works satisfactorily as a nearly unbiased point estimator when the abundances of rare species are nearly homogeneous or singletons and undetected species have approximately the same mean abundances (*23*). Because Chao1 is non-parametric, the lower bound is valid for any distribution of entities among types: it should be robust to differences in survival across DOCUMENT types (*15*).

Finally, we analyzed the evenness in these assemblages or the extent of equity among species abundances (*28*). A community's evenness will affect its stability in the face of external forcing, in particular its ability to withstand the impact of diversity-threatening events, such as wildfires (*29*). Given two equal-sized assemblages, the more even assemblage will be more resistant to the loss of WORKS through DOCUMENT losses. Below, we chart evenness profiles for one class (E_3) of evenness measures (Fig. 5). These curves can be connected to the slope of a Hill number profile: their steepness enables the intuitive comparison of the (un)evenness in the WORKS' abundances for the reconstructed assemblages (*21*).

Fig 3. Estimates for the union of the six assemblages. **A**: Hill number curves (for $0 \leq q \leq 3$), empirical and estimated, showing the absolute underestimation of the original diversity of WORKS. **B**: Species accumulation curve, plotting the number of WORKS as a function of the number of DOCUMENTS. The dot shows the observable data, the solid line the rarefaction for sample sizes $< n$, the dashed line the extrapolation to sample sizes $> n$. **C**: Kernel-density plot for the estimated number of DOCUMENTS.

The results for the union of the corpora (Table 1 and S2) suggest an overall survival ratio of 68.3% CI[63.2%–73.5%] for WORKS and 9.0% CI[7.5%–10.7%] for DOCUMENTS. The species accumulation curve (Fig. 3B) indicates at which rate we might still be discovering new WORKS in the future, by sighting more DOCUMENTS (*3*). Fig. 3A shows the empirical and estimated Hill number profiles: at $q = 0$ the curves indicate the absolute size of our current under-estimation of the original diversity in the combined assemblage of chivalric and heroic narratives from the medieval period. Of the original ∼1,170 WORKS that once would have existed, 799 would survive nowadays. Likewise, the 3,648 DOCUMENTS that are still observable constitute a sample from a population that originally would have counted ∼40,614 specimens (Fig. 3C).

We observe considerable inter-vernacular variation (Table 1), ranging from the relatively poorly surviving English WORKS (38.6%) to the relatively intact German tradition (79.0%). Dutch and French have a substantially lower survival factor than German, whereas two of the insular assemblages, Icelandic and Irish, have sustained similar losses to German, with point estimates of 77.3% / 81.0% and 16.9% / 19.2% for the survival of WORKS and DOCUMENTS respectively (*12*). Puzzling is that Old and Middle English documents did not travel far during their post-medieval afterlives (Fig. 4), while other literatures survive in a wide manuscript diaspora. The survival estimates for WORKS and DOCUMENTS yield similar rankings (Table 1). In the SM, we compare Chao1 to three other estimators with similar results (Fig. S1). Fig. 5, finally, shows the (estimated) evenness profiles and offers further insight into the distributional properties characterizing the assemblages. The profiles (Fig. S2) for additional evenness classes ($E_1 - E_5$) yield consistent findings. Here, too, we note the atypical nature of Icelandic and Irish: in comparison to the highly uneven distribution of e.g. French, these two insular literatures feature a much more even distribution of DOCUMENTS over WORKS.

Fig 4. Heatmap of the geolocations of the repositories where DOCUMENTS are kept for four vernaculars. Made with *Leaflet* 1.7.1.

Fig 5. Normalized evenness profiles (E_3) for the six individual vernaculars, plotting ${}^{q}E$ as a function of order $0 \leq q \leq 3$. The values on the Y-axis reflect the estimated evenness in the reconstructed assemblages.

language	f_1	f_2	S_{obs}	n	**Chao1**	**MS**
Dutch	45	13	75	167	0.492	0.075
English	42	8	69	176	0.386	0.049

language	f_1	f_2	S_{obs}	n	Chao1	MS
French	90	21	222	1473	0.535	0.054
German	36	19	128	1088	0.790	0.145
Icelandic	44	28	117	295	0.773	0.169
Irish	69	54	188	449	0.810	0.192
union	326	143	799	3648	0.683	0.090

Table 1. Point estimates of survival ratios in six traditions: for WORKS, using Chao1 (i.e., sample completeness at $q = 0$) and DOCUMENTS (MS) using the minimum sampling extension, including the number of works (S_{obs}), documents (n), singletons (f_1) and doubletons (f_2).

Regarding DOCUMENTS, our results confirm the severity of the losses, with survival ratio estimates ranging from 4.9% (English) to 19.2% (Irish). This corroborates previous estimates from book history, positing an overall survival factor of 7%, i.e. slightly lower than our point estimate for the union (9.0% CI[7.5%–10.7%]). Contrary to previous analyses (*16*, *17*), these results are therefore compatible with evidence from book history. It remains to be seen whether these estimates will scale to other cultural domains, but this analysis reveals important relative differences in the persistence of medieval heroic and chivalric narrative across Europe. Some of these differences have not been noticed before and challenge existing assumptions. For example, our results suggest that Irish and Icelandic literature have been preserved comparatively well compared to some of the more canonical mainland literatures (*12*).

In ecology, island ecosystems stand out: despite being comparatively species-poor for their land surface, they feature a higher endemic species richness compared to mainland regions (*30*). Additionally, insular assemblages demonstrate a higher species evenness, due to the lack of predators etc. A parallel emerges with some of the cultural diversity profiles for island regions reconstructed here: if land-isolated areas preserve biological heritage more effectively, the same might hold true for cultural heritage. Previous discussions about the survival of historic literature have focused on factors such as library fires or collectors' interests (*1*). We identify an additional key aspect that is typically overlooked: the evenness with which DOCUMENTS were originally distributed over WORKS fundamentally affected an assemblage's stability (*29*). Medieval French literature, for instance, was sizable, but its long tail of low-abundance works rendered it more susceptible to immaterial loss. Thus, while the loss figures for Icelandic and Irish are considerable, their distributional characteristics seem to have made them more resistant to post-medieval losses.

Which societies produce a highly even cultural output to safeguard the retention of their diversity? The role of demography, especially population size, has been hotly debated in cultural evolution (*6*, *7*, *31*). Smaller, isolated social groups can be more susceptible to the random loss of cultural traits because of stochastic drift (*6*), although these communities can adopt fitness-improving behavior to guard against such information loss. The topology of social networks seems crucial: a low network degree (or interconnectedness between individuals) can counter the impact of drift and promote the retention of cultural complexity (*32*). For the remote island of Rapa Nui, for example, a model-based account showed how structural constraints in social interactions might have stimulated the retention of diversity (*8*). We have extended these simulations (*21*) to show that a lower network degree, under neutral models of transmission, invariably leads to a more evenly distributed cultural production (Fig. S3).

References and Notes

1. E. Buringh, *Medieval Manuscript Production in the Latin West, Explorations with a Global Database* (Brill, 2011).

2. F. Bruni, A. Pettegree, Eds., *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (Brill, 2016).

3. N. J. Gotelli, R. K. Colwell, Estimating Species Richness, in *Biological Diversity: Frontiers in Measurement and Assessment* (Oxford University Press, 2011), pp. 39–54.

4. L. J. Gorenflo, S. Romaine, R. A. Mittermeier, K. Walker-Painemilla, Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*. **109**, 8032–8037 (2012).

5. H. Zhang, R. Mace, Cultural extinction in evolutionary perspective. *Evolutionary Human Sciences*. **3**, e30 (2021).

6. J. Henrich, Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses: The Tasmanian case. *American Antiquity*. **69**, 197–214 (2004).

7. J. Henrich, R. Boyd, M. Derex, M. A. Kline, A. Mesoudi, M. Muthukrishna, A. T. Powell, S. J. Shennan, M. G. Thomas, Understanding cumulative cultural evolution. *Proceedings of the National Academy of Sciences*. **113**, E6724–E6725 (2016).

8. C. P. Lipo, R. J. DiNapoli, M. E. Madsen, T. L. Hunt, Population structure drives cultural diversity in finite populations: A hypothesis for localized community patterns on Rapa Nui (Easter Island, Chile). *PLOS ONE*. **16**, 1–25 (2021).

9. A. Acerbi, J. Kendal, J. J. Tehrani, Cultural complexity and demography: The case of folktales. *Evolution and Human Behavior*. **38**, 474–480 (2017).

10. E. Kwakkel, *Books before Print* (Arc Humanities Press, 2018).

11. U. Neddermeyer, *Von der Handschrift zum gedruckten Buch. Schriftlichkeit und Leseinteresse im Mittelalter und in der frühen Neuzeit. Quantitative und qualitative Aspekte* (Harrassowitz, 1998).

12. D. Ó Corráin, What happened to Ireland's medieval manuscripts? *Peritia*. **22–23**, 191–223 (2011–2012).

13. R. Wilson, *The Lost Literature of Medieval England* (Methuen, 1970).

14. P. Eggert, *The Work and the Reader in Literary Studies: Scholarly Editing and Book History* (Cambridge University Press, 2019).

15. H. Wijsman, *Luxury Bound. Illustrated Manuscript Production and Noble and Princely Book Ownership in the Burgundian Netherlands (1400-1550)* (Brepols, 2010).

16. J. L. Cisne, How science survived: Medieval manuscripts' "demography" and classic texts' extinction. *Science*. **307**, 1305–1307 (2005).

17. G. Declercq, Comment on "How Science Survived: Medieval Manuscripts' 'demography' and Classic Texts' Extinction." *Science*. **310**, 1618–1618 (2005).

18. N. D. Pyenson, L. Pyenson, E. Buringh, J. L. Cisne, Treating medieval manuscripts as fossils. *Science*. **309**, 698–701 (2005).

19. M. S. Cuthbert, Tipping the iceberg: Missing Italian polyphony from the age of Schism. *Musica Disciplina*. **54**, 39–74 (2009).

20. L. Egghe, G. Proot, The estimation of the number of lost multi-copy documents: A new type of informetrics theory. *Journal of Informetrics*. **1**, 257–268 (2007).

21. Materials and methods are available as supplementary materials.

22. A. Chao, Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*. **11**, 265–270 (1984).

23. A. Chao, C. H. Chiu, Estimation of species richness and shared species richness, in *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, N. Balakrishnan, Ed. (Wiley, 2012), pp. 76–111.

24. M. I. Eren, A. Chao, W.-H. Hwang, R. K. Colwell, Estimating the richness of a population when the maximum number of classes is fixed: A nonparametric solution to an archaeological problem. *PLOS ONE*. **7**, 1–11 (2012).

25. A. Chao, Y. Kubota, D. Zelený, C.-H. Chiu, C.-F. Li, B. Kusumoto, M. Yasuhara, S. Thorn, C.-L. Wei, M. J. Costello, R. K. Colwell, Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*. **35**, 292–314 (2020).

26. M. O. Hill, Diversity and evenness: A unifying notation and its consequences. *Ecology*. **54**, 427–432 (1973).

27. A. Chao, R. K. Colwell, C.-W. Lin, N. J. Gotelli, Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*. **90**, 1125–1133 (2009).

28. A. Chao, C. Ricotta, Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*. **100**, e02852 (2019).

29. I. Donohue, H. Hillebrand, J. M. Montoya, O. L. Petchey, S. L. Pimm, M. S. Fowler, K. Healy, A. L. Jackson, M. Lurgi, D. McClean, N. E. O'Connor, E. J. O'Gorman, Q. Yang, Navigating the complexity of ecological stability. *Ecology Letters*. **19**, 1172–1185 (2016).

30. R. J. Whittaker, J. M. Fernández-Palacios, *Island Biogeography. Ecology, Evolution, and Conservation* (Oxford University Press, 2006).

31. A. Acerbi, R. A. Bentley, Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior*. **35**, 228–236 (2014).

32. M. Cantor, M. Chimento, S. Q. Smeele, P. He, D. Papageorgiou, L. M. Aplin, D. R. Farine, Social network architecture and the tempo of cumulative cultural evolution. *Proceedings of the Royal Society of London, Series B: Biological Sciences* (2021), doi:10.1101/2020.12.04.411934.

Acknowledgments: We would like to thank Dirk Schoenaers, Georgia Henley, Jean-Baptiste Camps, Bernd Bastert, Daniel Könitz, and Jeroen Deploige for their help, as well as the five anonymous referees.

Funding: EDB's work was funded through a postdoctoral fellowship from FWO Flanders; KAK's contribution was funded a Carlsberg Foundation Visiting Fellowship CF20-225 and a Carlsberg Foundation H. M. Queen Margrethe II Distinguished Research Fellowship CF18-500.

Author contributions: Conceptualization: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Data curation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Formal analysis: MK, FK. Investigation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Methodology: MK, FK, AC. Software: MK, FK. Validation: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Visualization:

MK, FK. Writing – original draft: MK, FK, EDB, MD, KAK, POM, DS, RS, AC. Writing – review & editing: MK, FK, EDB, MD, KAK, POM, DS, RS, AC.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: We share data and code in an open access repository [https://doi.org/10.5281/zenodo.4777804]. This paper is released with a Python software package (Copia, available from PyPI), all under a CC-BY-SA license, to replicate our findings and stimulate the adoption of this approach in other domains.

Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1 to S3

Tables S1 to S2

References (*33*–*98*)

Science AAAS

Supplementary Materials for

Forgotten Books: The Application of Unseen Species Models to the Survival of Culture

Mike Kestemont, Folgert Karsdorp, Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A. Kapitan, Pádraig Ó Macháin, Daniel Sawyer, Remco Sleiderink, Anne Chao

Correspondence to: mike.kestemont@uantwerp.be

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S3
Tables S1 to S2
References (*33*–*98*)

Materials and Methods

Methods

Species richness estimators
Beyond ecology, species richness estimation has found diverse applications, because the concept of "species" can be understood liberally. For examples, "species" can represent paradigmatic classes of stone tools in archaeology (*24*, *33*), die types in numismatics for ancient coins (*34*), bugs in a piece of software (*35*), organic pollutants discharged to a water environment (*36*), different cases for a specific disease in epidemiology (*37*), distinct operational taxonomic units (OTUs) in genomics (*38*), genes or alleles in genetics (*39*), vocabulary (distinct words) in linguistics (*40*), etc. An impressive number of species richness estimators have been proposed. Although we focus on a single popular estimation method (the Chao1 estimator) (*22*), three additional estimators, including the Jackknife procedure (*41*), the iChao1 estimator (*42*) and the Egghe and Proot estimator (*20*), are also presented for comparison below.

Based on the notation used in the main text, the Chao1 estimator (*22*) is expressed as $\hat{S}_{Chao1} = S_{obs} + \left(1 - (1/n)\right) f_1^2/(2f_2)$ if $f_2 > 0$. Here, n denotes the sample size, S_{obs} denotes the number of observed features; the number of undetected species is estimated by $\left(1 - (1/n)\right) f_1^2/(2f_2)$, which is a non-linear function of the counts of singletons and doubletons. In his renowned wartime cryptographic work, Alan Turing established that these two counts convey most of the available information about the undetected richness (*43*, *44*). Although an analytic formula exists for sampling variance and confidence interval of the Chao1 estimator (*45*), in this paper we adopt the bootstrap method detailed in the Appendix S2 to (*46*), mainly because the bootstrap method is applicable not only for the Chao1 estimator but also for all other estimators.

Historical note: The name "Chao1 estimator" for abundance data was given by Colwell and Coddington in an influential paper on species richness estimation (*47*). They also gave the name "Chao2 estimator" to a similar type of estimator based on replicated incidence or occurrence data. Recently, it was noted that the Chao1 point estimator (without assessing sampling uncertainties) was developed independently by the renowned astrophysicist Ambartsumian (who is often regarded as the founder of astrophysics in the Soviet Union) in his work on estimating the number of flare stars in the Pleiades (*48*).

General framework for the Chao1 estimator
The Chao1 estimator was derived as a lower bound of species richness under a very general framework: we assume there are an unknown number of classes or types (such as species, taxa, operational taxonomic units, genes, die types, categories, vocabularies, etc.) in an assemblage of entities (such as individuals, DNA sequences, coins, an author's words, etc.) Assume that a sample of entities are *independently selected or detected* from the assemblage and each entity can be correctly classified to its type. If data are not sparse so that frequencies of classes provide reliable information on singletons and doubletons, then the Chao1 method can be applied to any frequency count data and gives a universally valid nonparametric *lower bound* for the true number of classes. Here "non-parametric" means that the lower bound is valid for any class size distribution. From a theoretical perspective, for a hyper-diverse assemblage or severely under-sampled assemblage,

data of samples of limited size do not have sufficient information to accurately infer species richness due to insufficient information for rare species. Unless some strong assumptions are made, at best one can only evaluate lower bounds as there may have many undetectable or invisible species. A precise lower bound in many applications is preferable to non-accurate point estimates.

The Chao1 estimator is intuitive and very easy to calculate, and often performs as well as more complex asymptotic estimators (*3*). In addition, the Chao1 lower bound is robust in the following senses in an ecological context:

1. We reiterate that no assumptions are needed about the species abundance distribution. Some species may be very abundant and some species can be very rare. For example, in most assemblages, species abundances vary widely, with a few extremely abundant species and many rare species. The Chao1 method is valid for any species abundance distribution.
2. The underlying species abundance distribution can be relaxed to become a species detection-rate distribution. That is, we do not need to assume that all individuals or entities have the same probability of being selected in the sample, i.e., simple random sampling of individuals (or entities) is not necessary. Generally, species detection probability or rate in any observation is a combination of species abundance and individual detectability, which is determined by many possible factors (such as individual movement patterns, color, size and vocalizations). For example, we can assume that individuals of colorful species or large size have a higher rate of being detected without affecting the validity of the Chao1 method (*23*).
3. In vegetation surveys, species abundances of woody plants are often recorded in each of the selected plots or quadrats. Due to spatial aggregation or clumping, individual plants cannot be modeled as independent entities and thus the basic assumption that individuals are independently detected is not fulfilled. For example, the negative binomial distribution is typically used to describe spatially clustered (if sampling units are quadrats in an area) pattern of species frequencies. Under the negative binomial case, the Chao1 method is still valid as a lower bound (*42*). However, in our application, such “clumping” (e.g., of specific works in specific libraries) is exceedingly rare.

When is the Chao1 lower bound unbiased? A basic requirement for any proper species richness estimator is that it should be nearly unbiased when data are selected from a homogeneous assemblage. Here a homogeneous case means that all species have identical detection probability. The Chao1 estimator is nearly unbiased in the homogeneous case, and theoretically is a lower bound for non-homogeneous cases (*22*, *23*). Nevertheless, this does not imply that the Chao1 estimator cannot be unbiased in non-homogeneous cases. Based on the Good-Turing frequency formula derived from Alan Turing’s famous cryptographic work during WWII, Chao et al. (*44*) showed that if undetected species and singletons in a sample have identical mean detection probabilities, the Chao1 lower bound becomes an approximately unbiased point estimator. Therefore, a simple sufficient condition for the Chao1 lower bound to be nearly unbiased is that rare classes (specifically, singletons and undetected classes) have approximately homogenous detection probabilities; in this case, the abundant classes could be highly heterogeneous without affecting the nearly unbiasedness of the estimator.

Additional methods

In the paper, we focus on Chao1, which is an established and robust method in ecology. It is useful to compare the results, however, to additional estimators for the survival of works (*49*, *50*):

1. **Jackknife**: a statistical method for bias-correction (higher-order variant) (*41*)
2. **iChao1**: a variant of Chao1, that also considers f_3 (tripletons) and f_4 (quadrupletons) (*42*)
3. **Egghe & Proot**: estimator developed by Egghe and Proot (*20*), which was subsequently identified as a variant of an unseen species model (*51*) and which has found further adoption in early modern book history (*52*). Used with the default setting of alpha=150.

The survival ratios with respect to WORKS for the different sub-assemblages are included in Table S1 and visualized in Fig. S1. As can be seen, the methods show differences but are largely in agreement. The overall difference is confirmed between the relatively lower survival rates for Dutch, English and French, as opposed to the considerably higher survival rates for German, Icelandic and Irish. Note that the Jackknife and Egghe & Proot estimators may exhibit bias for the simplest homogeneous case, violating the basic requirement for a proper species-richness estimator. Simulation results revealed that Jackknife estimators often exhibit counter-intuitive patterns: their bias, accuracy and coverage probability of their confidence intervals regularly do not improve as sample size increases (*42*).

Minimum additional sampling

Based on an original sample of n individuals, the Chao1 estimator provides an estimate of the number of species not yet detected in the sample. Chao et al. (*27*) developed a method for estimating the minimum number of additional individuals required to discover all these undetected species. It follows from the Chao1 formula that there are no additional undetected species when every species is represented by at least two individuals (i.e., no singletons). In other words, sampling should continue until singletons vanish. They derived that the minimum required additional sample size is $m = nx^*$, where x^* satisfies the equation $2(1 + x^*)f_1 = exp[x^*(2f_2/f_1)]$. We estimate the number of DOCUMENTS in the original population as $n + m$. A bootstrap method (*46*) can be applied to obtain the sampling distribution of the minimum number of documents and the associated confidence intervals (see Fig. 3C in the main text).

Hill numbers and their estimators

Species richness does not take species abundance into account. Scientists in various disciplines developed “complexity” measures which consider both richness and abundance. For example, a popular complexity measure, the Gini-Simpson index, was originally developed by the Italian statistician Corrado Gini around 1912. The Gini-Simpson index represents the probability that two randomly chosen individuals belong to different species. Another popular measure of complexity is Shannon entropy, developed by Claude Shannon in 1948 in information science, which measures the uncertainty in the species identity of a randomly chosen individual in the assemblage. Although Shannon entropy (in units of information) and the Gini-Simpson index (a probability) have been widely used in various disciplines, they use different units and thus cannot be compared with each other. As indicated later, they can be converted via simple transformations to the same units of “species”.

Hill (*26*) integrated species richness and species abundances into a continuum of diversity measures, referred to as Hill numbers. Let p_i denote the relative abundance of species i. The Hill number of order $q \neq 1$ is defined as ${}^qD = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$, where the parameter q determines the measure’s sensitivity to species’ relative abundances. When $q = 0$, ${}^0D$ is simply species richness, which counts species equally without regard to their abundances. The Hill number of order $q = 1$ is defined as the limit of ${}^qD$ as q tends to 1; this limit is the exponential of Shannon entropy, referred to as "Shannon diversity". The measure for $q = 1$ counts individuals equally and can be interpreted as the number of abundant species in the assemblage. The Hill number of order $q = 2$ reduces to the inverse of the Simpson concentration index, referred to as "Simpson diversity", and can be interpreted as the effective number of abundant/common species in the assemblage.

Hill numbers for all orders $q \geq 0$ have the same units as species. They can therefore be depicted on a single graph as a function of q. This diversity profile conveys all the information in the species abundance distribution. In most applications, the empirical diversity profile generally underestimates, especially for $q \leq 1$. Chao and Jost (*46*) developed an analytic estimator of the profile to reduce the bias; they also derived a bootstrap method to assess the associated confidence bounds.

Evenness

Evenness quantifies the extent of equity amongst species abundances. In ecology, a wide range of evenness measures have been proposed under different criteria. A unified and intuitive concept of evenness that encompasses many of the most useful evenness measures is the following: unevenness (or its opposite, evenness) among species abundances is measured by a "distance" (or its opposite, "closeness") between the vector of species abundances and those of a completely even assemblage (*28*). Under this unified concept, Chao and Ricotta (*28*) derived five classes of evenness measures, all in terms of diversity ${}^qD$ (Hill number of order q) and species richness S. All the five classes (E_1–E_5, ignoring E_6, which cannot be applied to reconstructed assemblages) of evenness measures lie in a fixed interval [0, 1], with the minimum value of 0 for a maximally uneven assemblage (when one species is super-dominant, while all the others are vanishingly scarce), and the maximum value of 1 for a completely even assemblage (all species are equally abundant), regardless of species richness. This fixed range allows us to compare the evenness of assemblages with different numbers of species.

In the main text, we focus on the third class of evenness measures (E_3), which is expressed as ${}^qE_3 = ({}^qD - 1)/(S - 1)$ for $q > 0$ (*28*). (Note that for $q = 0$, species abundances are disregarded and evenness is not meaningful.) For sampling data, both the diversity ${}^qD$ and species richness S in the formula should be estimated via Chao and Jost's method (*46*); see Fig. 3A of the main text. One major reason that we focus on this evenness measure is that the corresponding unevenness measure can be connected to the slope of a diversity (Hill numbers) profile. From Fig. 3A in the main text, a diversity profile (which depicts diversity with respect to order q) is theoretically a non-increasing (constant or decreasing) function of order $q \geq 0$. When species abundances are even, the diversity profile is a horizontal line (slope = 0) at the level of species richness. Otherwise, the steepness of its slope reflects the unevenness of species abundances. When species richness is fixed, the more uneven species abundances are, the more steeply the profile decreases. In Fig. 5 of the main text, we quantify evenness for each vernacular through a continuous profile that depicts evenness as a function of diversity order $q > 0$. The profiles can be visually compared across the six assemblages, even if they have different numbers of species (*25*).

Evenness and Stability

The evenness among species abundances is an important distributional characteristic of an assemblage, which ties into the concept of 'ecological stability' (*29*, *53*). Renowned ecologist MacArthur (*54*) was the first to use an evenness measure (Shannon entropy) to quantify community stability in a trophic web. Hairston et al. (*55*) proposed that community stability was measured as a tendency to maintain evenness of the species abundance distribution; they also adopted an evenness measure based on Shannon entropy. An important component of ecosystem stability is (temporal) invariability, which has been measured by the degree of evenness of community-aggregate biomass, productivity or other pertinent measures (*53*, *56*–*58*). This is an evenness measure based on the inverse of coefficient of variation (CV, the ratio of mean to standard deviation).

All these stability/invariability measures based on Shannon entropy and CV can be integrated into a class of measures also derived under Chao and Ricotta's (*28*) unified concept of evenness. This class of evenness measures is expressed as ${}^{q}E_1 = [1 - ({}^{q}D)^{1-q}]/(1 - S^{1-q})$, i.e., the first class in (*28*). When q tends to 1, it reduces to Shannon-entropy divided by $\log(S)$, which is known as Pielou's J' evenness index (*59*); for $q = 2$, the corresponding evenness measure is $1 - CV^2/S$. In Fig. S2, we show the estimated evenness profiles for this class of evenness measures along with the other classes based on our data (E_1–E_5). The plots for five classes of evenness measures exhibit consistent patterns among the six vernaculars in that Icelandic and Irish reveal comparatively higher values of evenness.

Simulations: evenness and network structure

In the domain of cultural evolution, the relationship between demography and (cumulative) cultural complexity has been hotly debated (*7*, *9*). Research in this area commonly makes use of mathematical simulation models to infer which processes of social learning and transmission can potentially explain the observed frequencies of cultural data that survive from the past (*60*). Regarding demography, it has been convincingly argued that, because of the sampling error associated with random drift, small and/or isolated communities are in principle more susceptible to the loss of cultural traits (*6*, *61*). A variety of studies have established that the topology of a society's network structure is able to counter the effects or drift: a lower degree of interconnectedness between individuals (and subcommunities) will stimulate the production and retention of cultural diversity (*8*, *32*, *62*).

To specifically investigate the aspect of evenness in cultural production, we build upon the analyses reported by Lipo et al. (*8*) for the case study of Rapa Nui. Here, we employ a simulation model to study the relation between network degree and evenness. The model is an extension of the Wright-Fisher model (*63*), which assumes a population of constant size N and discrete, non-overlapping generations. The extension adds a network component to the standard Wright-Fisher model, allowing us to control the number of connections k between individuals. When $k = N - 1$, the model reduces to the standard Wright-Fisher model, in which the sample pool of each individual consists of the entire population. By contrast, with $k = 2$, the sample pool of each individual is limited to its immediate neighbors. Assuming a constant degree k, then, the probability of selecting some cultural variant i by an individual is equal to:

$$p_i = n_i/k(1 - \mu)$$

where n_i refers to the number of occurrences of variant i in the individual's sample pool of size k. Individuals faithfully copy a cultural variant with probability $(1 - \mu)$ and innovate a new one with probability μ. As in the classical Wright-Fisher model, generation transition happens simultaneously. We only experiment with neutral, unbiased transmission in order to narrow down the effect of network structure on evenness and leave other, biased forms of transmission for further research (*31*, *60*, *64*).

We simulate populations with the following parameter ranges: $N \in \{10^3, 5 \times 10^3, 10^4\}$, $\mu = 0.005$ and $k \in \{2,5,10,20,100\}$. Populations are initialized with $T = 2$ distinct traits. Once these initial traits are lost, we can assume the cultural system has reached its equilibrium state. We then simulate an additional 10^4 generations and compute the evenness profile E_3 for the cultural production of the final population, specifically for the E_3 class of evenness measures. The results presented in Fig. S3 display a clear effect of network degree on the evenness profiles. The graphs depict the mean evenness profiles computed based on 100 simulations. For all population sizes, low-degree networks consistently yield assemblages with higher evenness values (equity in species abundances), whereas high connectivity between individuals results in lower E_3 values.

We refrain from comparing our simulations to empirical data because of various reasons. The population sizes for medieval geopolitical entities are hard to estimate (*65*), as are their *effective* population sizes, because (active nor passive) literacy rates are notoriously intractable for the medieval period (*66*). Literacy in a particular language doesn't map, in this period, onto territorial boundaries: e.g., literacy in French extended through the upper echelons of society well beyond the borders of the kingdom of France and so on. Moreover: much literature was read aloud (in group), instead of being read individually or silently (*66*).

Materials

Medieval chivalric and heroic narrative fiction

The body of medieval vernacular heroic and chivalric fiction includes numerous works (*13*, *67–69*), with dates of composition ranging across the entire medieval millennium (*70*). Subsequent generations of authors frequently reworked such chivalric or heroic narratives. These stories also frequently travelled across multiple linguistic communities through processes of (intervernacular) translation and adaptation. Below, we justify the composition and provenance of each subcorpus (or "assemblage") in the data underpinning the paper. Our analysis has been restricted to longer-form, vernacular, narrative fiction from the Middle Ages (ca. 600–1450 CE (*70*)) that is concerned with heroic and chivalric settings, themes and protagonists. Prose as well as verse texts (e.g. rhymed, alliterative, assonantic) were included. No distinction was made between translated, adapted and original materials. All texts belong to the narrative or "epic" genre, in the sense that "dramatic" and "lyrical" texts were explicitly barred. We excluded historiography (e.g. chronicles), hagiography (saints' lives), allegorical literature (e.g. *Le Roman de la Rose*) and animal fables (e.g. *Van den vos Reynaerde*). The main subvarieties in the data include (i) Arthurian romances, (ii) *chansons de geste* and other literature of the heroic type, (iii) fiction inspired by Classical antiquity and (iv) crusader fiction. Apart from the conventional "courtly romance", we also considered heroic narratives (e.g., the epic *Beowulf*, the Icelandic *fornaldarsögur* or the Irish tales).

In this study, the WORK and the DOCUMENT are the primary units of analysis (*14*). The attestation of a WORK in a (handwritten) DOCUMENT (also called 'text carrier', 'textual witness' or, simply, 'copy') is treated analogously to the "sighting" of a species in an ecological sampling campaign. The identification (discretization) of individual WORKS (in the sense of a biological species) is an interpretative act which was outsourced as much as possible to authoritative, pre-existing resources (either printed repertories or digital databases). Thus, for our purposes, a WORK is a literary concept which existing scholarly surveys have treated as unitary, despite any textual variation between the DOCUMENTS transmitting it. The concept of the discrete, non-material WORK represented by one or more surviving textual documents is widely shared among literary and textual scholars across different schools of textual-critical thought (*14*, *71*, *72*). Medieval documents can vary textually across different documents (*67*, *73*, *74*), but such variation does not, however, prevent the identification of one discrete work (*75*). Below, we concisely describe the process of collecting and preprocessing the data, vernacular by vernacular. All materials (code and data) for replicating our results have been publicly deposited (DOI:10.5281/zenodo.4777804), under a CC-BY-SA license, encouraging their re-use.

For each document in the dataset, we minimally record its (i) current shelfmark, (ii) current repository and (iii) the standardized title of the WORK which it represents. If retrievable, we have semi-automatically added an (iv) approximate geolocalization of the current location of a document via longitude-latitude coordinate pairs. The recorded attestations are limited to medieval documents that bear textual witness to medieval works. Only handwritten documents were included, regardless of their survival status or completeness (e.g., intact codices vs. binding waste); printed documents were excluded. The allowed dates (for works and documents) were limited to the medieval period, from the very start of the vernacular traditions involved up to the introduction and widespread adoption of the printing press in the respective linguistic communities and geographies (a process that began ca. 1450 in some parts of Europe, but took time and did not happen at the same moment in different places). An individual document might appear multiple times in the data (with the same signature), if it contains multiple works (miscellanies and anthologies). If a single manuscript survives in disjoint fragments (across different locales), the fragments were grouped into a single attestation.

Vernacular sub-assemblages included

Dutch (*Middelnederlands*)

The collection of medieval Dutch chivalric romances (*ridderepiek*) (*76*) has been listed in an authoritative repertory (*77*). The dataset was manually extracted from this resource and supplemented with more recently discovered data (*49*). Prose texts are rare in this literature, but were included nonetheless. The materials included translations, adaptations and original works from the main classical subvarieties of chivalric fiction: Arthurian romances (*Arturepiek*), chansons de geste (*Karelepiek*), crusader narratives (*kruisvaartepiek*) and antiquity-inspired fiction (*antikiserende roman*). Previous studies have surveyed the chronology and geography of the corpus (*78–80*).

French (*langues d'oïl; ancien et moyen français*)

For French literature, the longer-form narrative chivalric fiction is primarily represented by the subgenres of the *chanson de geste* and the *roman*, existing both in verse and prose (*81*). These works are listed as chansons de geste, romans français and romans idylliques in the online repertory Arlima (*Archives de littérature du Moyen Age*). (Note, however, the different listing of the cyclic texts.) Works in Occitan (*langue d'oc*), Anglo-Norman and Franco-Italian were excluded, as well as all shorter narratives (*fabliaux* and *dits*) and textual traditions related to *Roman de Renart* and *Roman de la Rose*. Works and documents dated after ca. 1450 were excluded. For the metadata on the manuscripts, we made use of IRHT's online repertory Jonas (*Répertoire des textes et des* *manuscrits médiévaux d'oc et d'oïl*): all works retrieved from Arlima were manually disambiguated and linked to oeuvres in Jonas. This combination was necessary, because Jonas offers more complete, recent and better accessible information on the material sources, but does not offer Arlima's option to select works on the basis of genre. In compiling the information from the two websites, executive decisions had to be made as to whether two texts were to be considered as different works (two different 'species') or as two *rédactions* of the same text. Similarly, we had to make pragmatic choices as to whether we considered different branches as part of one cyclic text (e.g. the different constituents of *Le cycle de la Croisade*), or as texts in their own right (e.g. *Enfances Godefroi*). In line with the research tradition and the way these texts often were composed and were circulating (both individually and in cyclic form), we typically treated the larger parts of a cycle as different works. We consulted the scholarly literature in individual cases where Arlima and Jonas were conflicting or were insufficiently informative (e.g. for the date of the composition of a work or the date of a manuscript).

High German (*Mittelhochdeutsch*)

In Middle High German studies, 'chivalric romances' are not commonly studied as a distinct genre. The categories used are *Romane* (romances), including Arthurian romance, classical romance and *Minne- und Aventiureromane* (romances on love and adventure) and *Heldenepik*, epic texts featuring (Germanic) heroes. Some studies categorize *chansons de geste* as a subgroup of *Heldenepik*, others consider them to be a separate genre. Since a database or bibliography including both romances and epic texts does not exist in German studies, the German dataset was derived from authoritative overviews on *chansons de geste* (*82*), 'Heldenepik' (*83*), classical romance (*83*) and Arthurian romance (*84*), as well as from the renowned literary history of Brunner (editions from 1997 onwards) (*85*). Subsequently, the manuscript data were extracted from the online database Handschriftencensus, after each page was manually linked to the correct work ID in this particular resource.

English (*Old and Middle English*)

For the Old English period (up to ca. 1150), most of the surviving heroic works were too short to qualify (*86*). We included the two longer-form works *Beowulf* and *Waldere* but excluded the *Battle of Brunanburh* and the *Five Boroughs* (because they were too short to qualify as a 'romance') and the *Finnsburh Fragment* and *The Battle of Maldon* because the original text carriers do not survive. *Beowulf* has had a disproportionate impact on popular perceptions of Old English verse: it is the only sustained heroic Old English narrative surviving in a near-complete state. *Waldere* survives only in two short fragments from one copy, but comparison to analogues and the level of detail in the fragments strongly suggest a sustained work, probably exceeding a thousand lines. A similar lack of known examples prevails in 12th and 13th-century Middle English. There is, though, the ca. 16,000-line-long, partly-Arthurian *Brut* of Layamon, which we include. Layamon's emphasis on violence and heroism rather than chronicling or historical recounting distinguishes his poem from the many versions of the later Middle English Prose *Brut* and places it within our study's purview.

Later Middle English studies conventionally uses 'romance' as a baggy umbrella term for adventurous or marvellous stories, including those ostensibly based on events regarded at the time as history, such as the three "matters" of Britain, France and Rome. Compared to surviving texts in German and French, the corpus of Middle English romances is limited in size. Writers composed them in both prose and verse, but a large majority of texts in the surviving corpus are in verse. Unlike in some other contexts, such as Iceland, in England print began to affect the circulation of vernacular romances soon after its first uses for transmitting any English at all. This gives a cut-off date of ca. 1475. The current standard survey is the [Database of Middle English Romance](). Manuscript transmission records were extracted from this database and then pruned to eliminate witnesses copied after the cut-off date. Romances only known today in copies later than the cut-off date were removed entirely. The *Tale of Gamelyn* survives in a striking number of copies because it was part of the textual tradition of Geoffrey Chaucer's *Canterbury Tales* and travelled within copies of Chaucer's text. *Gamelyn* is probably not itself Chaucer's work and so we regard the surviving copies in *Tales* manuscripts as legitimate, if odd, instances of a romance's survival.

Anglo-Norman

After the conquest of England by the Normans in 1066, French long held a place as an alternative, more prestigious vernacular, with particular purchase in government, the nobility and the legal system. French as spoken in England developed from the Norman variety into a distinct type, variously called in modern scholarship 'Anglo-Norman' (especially in its earlier phases), 'Insular French' (especially in its later phases), or 'the French of England' (*87*). Romances written in Anglo-Norman have attracted fewer studies, suffering because they excite neither scholars of Middle English nor scholars of continental French. However, the evidence suggests that they performed the same functions as Middle English romance, for the same audiences, in the same places; a striking number of Middle English romances adapt Anglo-Norman predecessors. There is therefore a strong, long-standing case (*88*) for regarding Anglo-Norman and Middle English romances as part of one tradition which happened, as the linguistic landscape of its audience changed, to switch languages partway through its history. The corpus of surviving Anglo-Norman romances is small, too small for our quantitative methods to be applied reliably to it on its own. We have, though, considered the Middle English romances alone and then the Middle English romances together with the Anglo-Norman romances as one insular group (see below).

We gathered texts and manuscripts from Ruth J. Dean's survey of Anglo-Norman texts and their manuscript witnesses (*89*), the current standard authority. As with the English material, some of the Anglo-Norman texts require judgments: genre taxonomy challenges any reference tool and Dean's excellent work is no exception. We therefore did not limit ourselves to the texts she lists as romances, but examined her other sections too, eventually including a few further texts which are sufficiently violent, adventurous and heroic to qualify. Often Dean herself comments on the multiple generic affiliations of these. A good example of such a text is *William the Marshal*, which is, in a sense, a biography of a real historical figure (William Marshal, d. 1219), but treats his life in thoroughly heroic terms. Wace's early Anglo-Norman *Roman de Brut*, the source for Layamon's *Brut*, is less violent and less heroic than its Early Middle English descendant, but we nevertheless include it, as it treats King Arthur at length and displays more of the interests of romance than its own source, Geoffrey of Monmouth's *Historia regum Britanniae*.

Icelandic (Old Norse-Icelandic, (*íslenska*; *norræna*)

Sagas

Several popular genres of medieval Old Norse-Icelandic literature fall under the term longer-form, vernacular narrative fiction. The material that most closely corresponds to this genre definition are the *riddarasögur* (lit. "sagas of knights"). These are essentially chivalric romances, either 13thcentury translations of predominantly French *romans courtois*, or younger original Icelandic compositions using some of the same characters, settings and motifs as the translated romances. There are about a dozen of the former, most of them translated in Norway but found almost exclusively in Icelandic manuscripts. The latter, the locally generated romances, number between ca. 30–40. A vast number – well over 200 – survive from the post-medieval period, but are deliberately excluded from the present data. The *fornaldarsögur* (lit. "sagas of ancient times") are also very much in the romance vein – formulaic, episodic quest narratives with a strong element of the fabulous and a happy ending – but take place in a Viking rather than a chivalric milieu and deal with the early history of Scandinavia – "la matière du nord" (*90*). There are about 35 of these sagas preserved from the medieval period and roughly an equal number composed after the Middle Ages, which, like the post-medieval *riddarasögur*, are not included in this study. In literary histories it is common to distinguish between these different types, even though they all have more in common with each other than not, and there are many borderline cases. As is clear from their transmission, early readers do not appear to have made a distinction between the two, as they are generally found alongside each other in the manuscripts. All of this material remained immensely popular in Iceland, well into modern times, and it is not unusual to find sagas preserved in 50, 60 or 70 manuscripts, as well as in metrical versions, known as *rímur* (lit. "rhymes", see below).

For the present study, information was gathered on all known manuscripts in which texts of *fornaladarsögur* and *riddarasögur*, both translated and locally generated, are preserved. As mentioned, we have only included works demonstrably from the medieval period, i.e. which survive in medieval manuscripts, and ignored those which are only preserved in younger manuscripts. Even for these medieval works, the majority of surviving witnesses are generally post-medieval, but for the sake of comparison with other traditions we have excluded them. The cut-off date for medieval/post-medieval is by common consent taken to be the mid-16th century (1550 being the date of the completion of the Protestant Reformation in Iceland). This also happens to be the period in which books began to be printed in Iceland, but it should be noted that the advent of print had almost no effect on manuscript production in Iceland, which continued unabated well into the modern era. The information on the *riddarasögur* is based on (*91*), which has been updated largely based on the printed supplements to the catalogue of the National and University Library of Iceland (*92*), and more recently online. Only information on the prose sagas is given in Kalinke and Mitchell's bibliography (*91*), so a list of manuscripts containing medieval *rímur* based on *riddarasögur* had to be compiled separately (see below) (*93*). Information on the *fornaldarsögur* is taken from the database of the *Stories for all time* project, which contains descriptions of all the manuscripts in which *fornaldarsaga* texts are found. Again, for the purposes of the present study account was taken only of the medieval witnesses.

Rímur

Many Old Norse-Icelandic sagas have counterparts in the metrical form of *rímur*, which started to appear in the fourteenth century and remained a popular literary form in Iceland until the modern era. Due to their close relationship to the prose versions of the stories that they are narrating, *rímur* related to the legendary sagas (*fornaldarsögur*) and chivalric sagas (*riddarasögur*) were also included in our experiments. While no database of *rímur* yet exists, the data was compiled manually based on authoritative printed publications (*92–96*) and online catalogues of Nordic manuscripts, such as the Handrit, the Dictionary of Old Norse Prose Register and the Skaldic Project.

Irish (Middle and Early Modern Irish/Gaelic)

The Gaelic medieval narrative tradition comprises of different thematic genres that include, for example, origin legends, voyage tales (*immrama*), wooing tales (*tochmairc*), conception tales (*comperta*), siege-tales (*togla*), cattle-drives (*tána*) and death tales (*aideda*); together with a wider classification of adventures (*echtrai*), and the extensive saga literature particular to the 'Ulster Cycle', which has at its centre-piece the great tale *Táin Bó Cuailnge*. It also includes narratives associated with kingship and individual kings of Ireland, and the most enduring of all genres, that of *fianaigheacht*, which has as its focus the deeds of Fionn mac Cumhaill and his band (*fian*) of warriors. Collectively, this narrative tradition formed an important component of what was termed *seanchas* ("learned tradition"), which also encompassed genealogical, historical, legal, medical, hagiographical and religious writings in verse and prose, and texts based on or translated from the Classical traditions. *Seanchas* texts form the subject matter of many of the great vernacular Gaelic manuscripts of the late Middle Ages.

Some of the narrative literature intersects structurally and thematically with Classical literature and with romance and chivalric literature of non-Gaelic traditions (episodic plots, quests, battles, otherworld encounters etc.) but this is largely coincidental. European literature in Gaelic tradition takes the form of translations and adaptations. In the case of Arthurian tales, works in translation (the Grail legend, for example) survive from late-medieval times, and there is also a large number of romance-inspired tales composed in the Modern-Irish period (17th and 18th centuries). For the purposes of the present study, the late romance material is not included. The local, Gaelic texts in the dataset reflect those classified by Ó Corráin (*97*) as medieval vernacular narrative prose. The cut-off point for the inclusion of witnesses is 1600, the end of the vellum era in Ireland. This terminus has necessitated the exclusion of the following medieval and early-modern texts that survive in late witnesses only (a feature of Gaelic textual tradition): *Buile Shuibhne*, *Caithréim Chonghail Chláiringhigh*, *Cath Cumair*, *Cath Findchorad*, *Cath Gabhra*, *Eachtra Chonaill Ghulban*, *Foghlaim Chon Culainn* and *Tóruigheacht Dhiarmada agus Gráinne*. Of particular relevance in this context is a study by D. Ó Corráin (*12*): 'Scholars do not tend to appreciate how slender the transmission is. Of 117 Old- and Middle-Irish narrative texts (and that includes nearly all), 34 survive in a single manuscript (and at least three of these are textually defective); 31 survive in only two manuscripts; and only 52 in three manuscripts or more.'

Supplementary Text

Factors of survival

A number of, primarily post-medieval, factors have been identified as playing a major role in the survival of books, apart from their individual (im)material value. The early accession of books in institutional or family libraries greatly enhanced their survival chances (*15*), although historical events, such as (Counter-)Reformation iconoclasm or bombing during WWII, led to the destruction or dissolution of many such sanctuaries. There are, moreover, considerable differences in the library histories of the areas considered here. For instance, while Ireland can boast some of the earliest recorded (vernacular) literature in Europe, the word for 'library' in Gaelic was a 17th-century lexicographer's invention that did not gain currency until the 19th century. The interests of (early) modern collectors also co-determined the shape of survival (*98*). These interests often both flowed from and contributed to the construction of national identities in the romantic period, and the assignment of complex medieval pasts to unitary modern nation-states in the nineteenth and twentieth centuries.

Secondly, the introduction of movable-type printing, tightly associated with processes of urbanization, is a major factor. Organized local manuscript production could quickly wane after the emergence of printing, and the availability of printed books containing the same works as existing manuscripts could make the preservation of those manuscripts seem less important. At the same time, however, many medieval works have only survived (albeit fragmentarily) because major printing centres in e.g. the Low Countries or Germany were in constant need of binding waste. Here too, striking geographic differences exist: Iceland, for instance, only had a single printing press from about 1530 up to the 18th century and manuscript book production continued there alongside printing well into the modern era. In Ireland, printing arrived even later (in 1551) and for long remained the preserve of (Counter-)Reformation works, with a negligible effect on the manuscript tradition.

Caveat: language ≠ geography

The survival estimates for the Old and Middle English corpus, rich in singletons, are remarkably low (4.9% for DOCUMENTS and 38.6% for WORKS). These results, at least for the present dataset, are more pronounced than those in previous research, which had already observed a lower survival rate for English literature than for its continental counterparts, but which remained statistically inconclusive (*1*). An important caveat is that the present approach is based on language rather than geography and thus ignores literary production in England in languages other than English. It is therefore useful to consider the corpus of heroic narratives in "Anglo-Norman", a type of Old Norman French, written and circulated in Britain from the conquest of England by the Normans in 1066 into the 15th century (*89*). The Anglo-Norman corpus is centred in England and written in a type of French at once distinctive from yet also clearly related to continental French; as a geographic hybrid, it resists easy classification. There is a long-standing case for treating Anglo-Norman and Middle English romance as part of a single tradition (*88*). We therefore compiled abundance data for Anglo-Norman literature and added it to the English data to assess the effect of its exclusion. The Anglo-Norman corpus is well-preserved and ups the survival rate for English WORKS considerably (51.6% CI[36.7% – 67.9%]), although the resulting combined rate remains far from the highest survival rates observed here. This inclusion somewhat normalizes the English situation, but does not resolve the question of why literature in English itself seemingly survived so poorly.

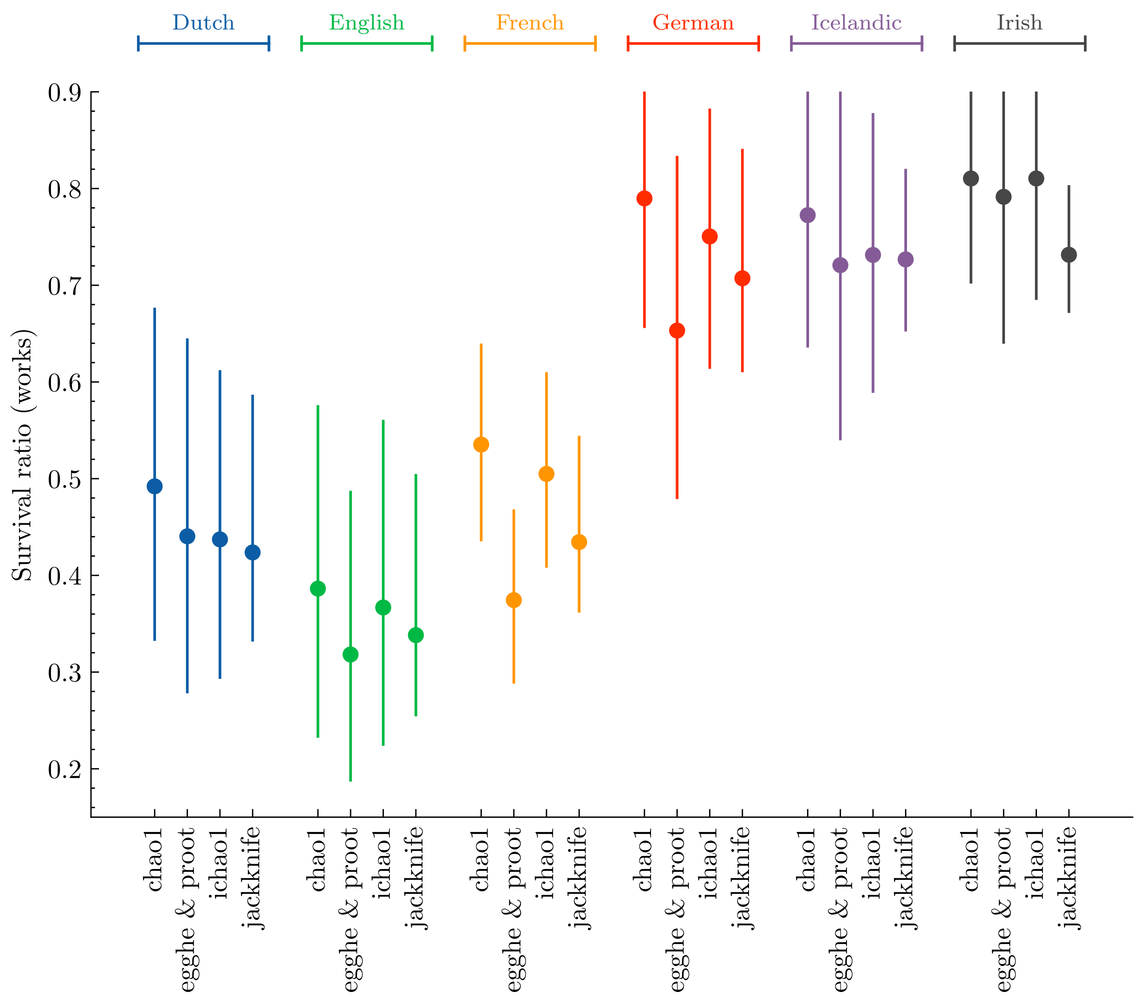


Fig. S1.

Error bar plot, showing the additional survival ratio estimates for WORKS in the six sub-assemblages, for three alternative unseen species methods. Shown are the point estimates for each vernacular–method combination, including the lower and upper confidence interval resulting from the bootstrap procedure.

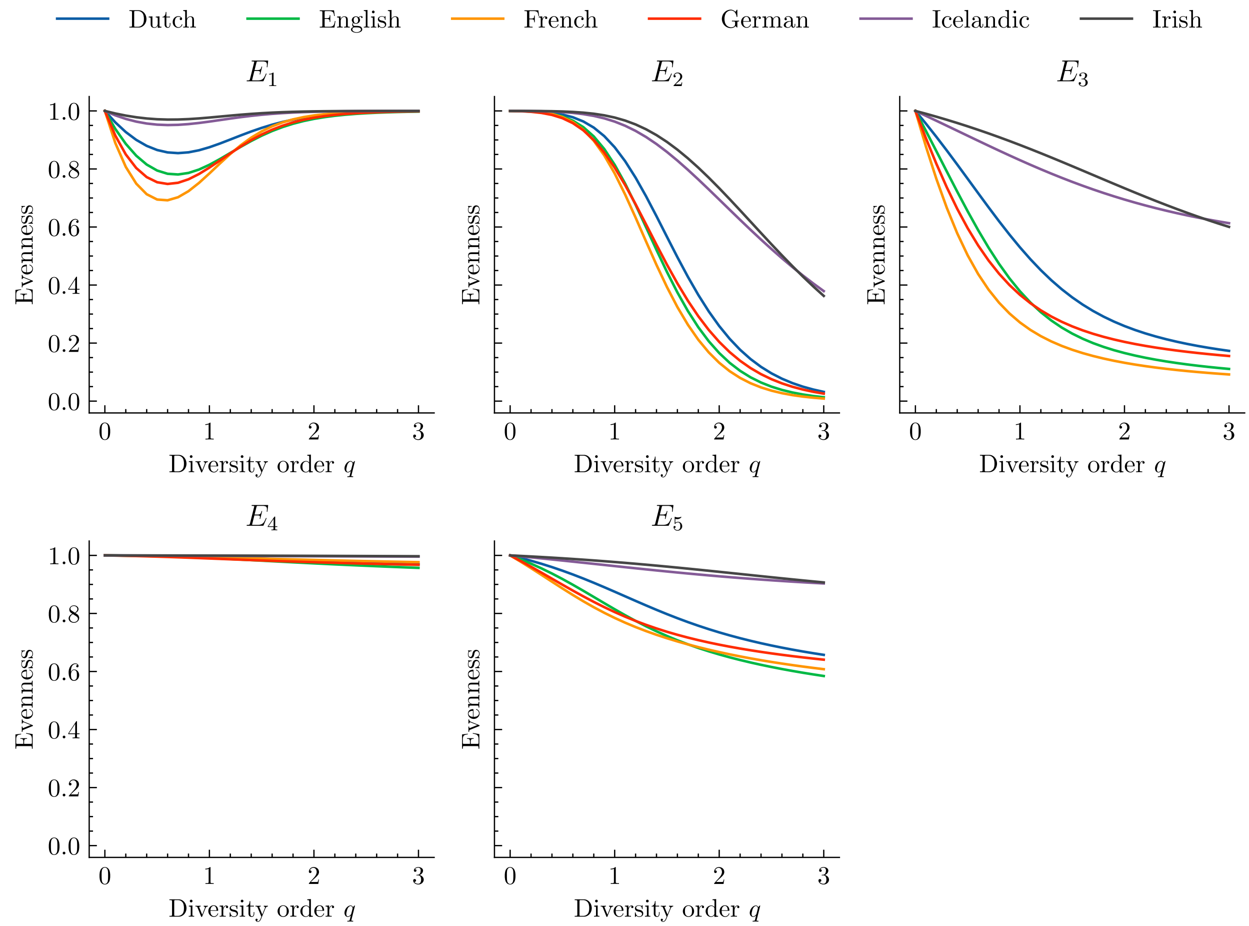


Fig. S2.

Evenness profiles for all six assemblages, for each class (E_1-E_5) of evenness measures distinguished by Chao & Ricotta (*28*). The profile for E_3 is identical to Fig. 5 in the main paper.

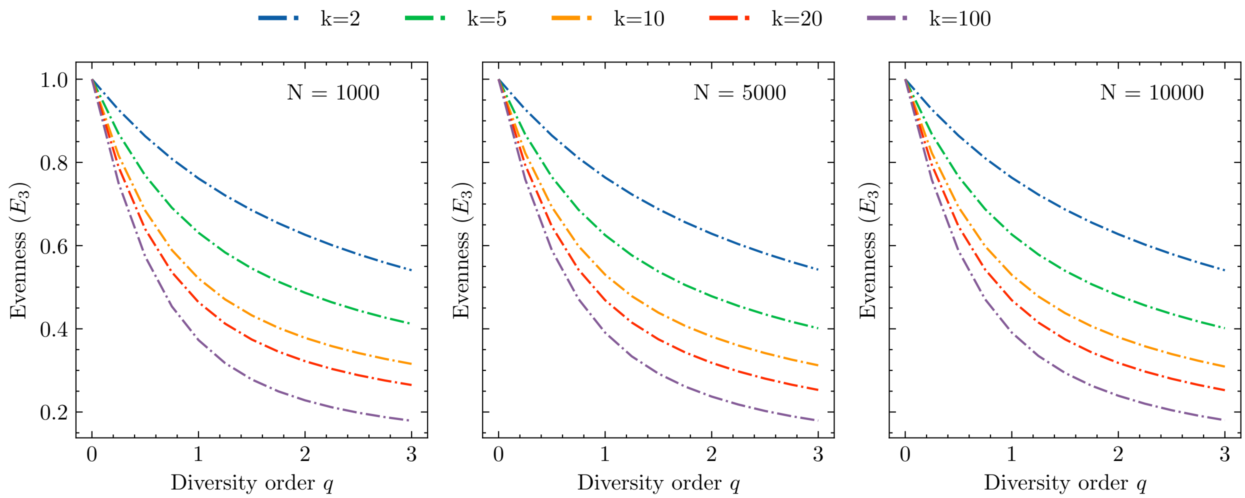


Fig. S3.

Simulations showing the effect of an artificial assemblage's social network structure on the evenness of its cultural production. For different population sizes ($N \in \{10^3, 5 \times 10^3, 10^4\}$) and 100 simulations, we show the mean effect of the network degree k *(*$k = 2,5,10,20,100$*)* on the evenness profile for the populations' cultural production, specifically for the E_3 class of evenness measures.

tradition	estimator	survival	lci	uci
Dutch	Chao1	0.492	0.335	0.679
Dutch	Egghe & Proot	0.440	0.282	0.644
Dutch	iChao1	0.437	0.294	0.612
Dutch	Jackknife	0.424	0.332	0.587
English	Chao1	0.386	0.236	0.572
English	Egghe & Proot	0.318	0.185	0.490
English	iChao1	0.367	0.221	0.560
English	Jackknife	0.338	0.254	0.505
French	Chao1	0.535	0.434	0.640
French	Egghe & Proot	0.374	0.287	0.469
French	iChao1	0.505	0.407	0.608
French	Jackknife	0.434	0.361	0.544
German	Chao1	0.790	0.654	0.918
German	Egghe & Proot	0.653	0.475	0.838
German	iChao1	0.750	0.613	0.884
German	Jackknife	0.707	0.610	0.841
Icelandic	Chao1	0.773	0.633	0.908
Icelandic	Egghe & Proot	0.721	0.541	0.911
Icelandic	iChao1	0.731	0.593	0.879
Icelandic	Jackknife	0.727	0.652	0.820
Irish	Chao1	0.810	0.703	0.918
Irish	Egghe & Proot	0.791	0.646	0.946
Irish	iChao1	0.810	0.684	0.946
Irish	Jackknife	0.732	0.671	0.804

Table S1.

Tabular representation of the additional survival ratio estimates for WORKS in the six sub-assemblages, for three alternative unseen species methods. Shown are the point estimates for each vernacular–method combination, including the lower and upper confidence interval resulting from the bootstrap procedure.

language	f_1	f_2	S_{obs}	n	repo	Chao1	Chao1-lCI	Chao1-uCI	MS	MS-lCI	MS-uCI
Dutch	45	13	75	167	80	0.492	0.339	0.666	0.075	0.040	0.140
English	42	8	69	176	36	0.386	0.229	0.570	0.049	0.024	0.087
french	90	21	222	1473	234	0.535	0.431	0.642	0.054	0.038	0.074
German	36	19	128	1088	389	0.790	0.648	0.918	0.145	0.086	0.255
Icelandic	44	28	117	295	12	0.773	0.632	0.905	0.169	0.102	0.286
Irish	69	54	188	449	15	0.810	0.703	0.911	0.192	0.119	0.310
union	326	143	799	3648	*NA*	0.683	0.632	0.735	0.090	0.075	0.107

Table S2.

Expanded version of Table 1 in the main paper, including upper and lower confidence intervals (lCI and uCI) for the Chao1 and MS-estimates, as well as the number of unique repositories for each assemblage.

References and Notes

1. E. Buringh, *Medieval Manuscript Production in the Latin West, Explorations with a Global Database* (Brill, 2011).
2. F. Bruni, A. Pettegree, Eds., *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (Brill, 2016).
3. N. J. Gotelli, R. K. Colwell, Estimating Species Richness, in *Biological Diversity: Frontiers in Measurement and Assessment* (Oxford University Press, 2011), pp. 39–54.
4. L. J. Gorenflo, S. Romaine, R. A. Mittermeier, K. Walker-Painemilla, Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*. **109**, 8032–8037 (2012).
5. H. Zhang, R. Mace, Cultural extinction in evolutionary perspective. *Evolutionary Human Sciences*. **3**, e30 (2021).
6. J. Henrich, Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses: The Tasmanian case. *American Antiquity*. **69**, 197–214 (2004).
7. J. Henrich, R. Boyd, M. Derex, M. A. Kline, A. Mesoudi, M. Muthukrishna, A. T. Powell, S. J. Shennan, M. G. Thomas, Understanding cumulative cultural evolution. *Proceedings of the National Academy of Sciences*. **113**, E6724–E6725 (2016).
8. C. P. Lipo, R. J. DiNapoli, M. E. Madsen, T. L. Hunt, Population structure drives cultural diversity in finite populations: A hypothesis for localized community patterns on Rapa Nui (Easter Island, Chile). *PLOS ONE*. **16**, 1–25 (2021).
9. A. Acerbi, J. Kendal, J. J. Tehrani, Cultural complexity and demography: The case of folktales. *Evolution and Human Behavior*. **38**, 474–480 (2017).
10. E. Kwakkel, *Books before Print* (Arc Humanities Press, 2018).
11. U. Neddermeyer, *Von der Handschrift zum gedruckten Buch. Schriftlichkeit und Leseinteresse im Mittelalter und in der frühen Neuzeit. Quantitative und qualitative Aspekte* (Harrassowitz, 1998).

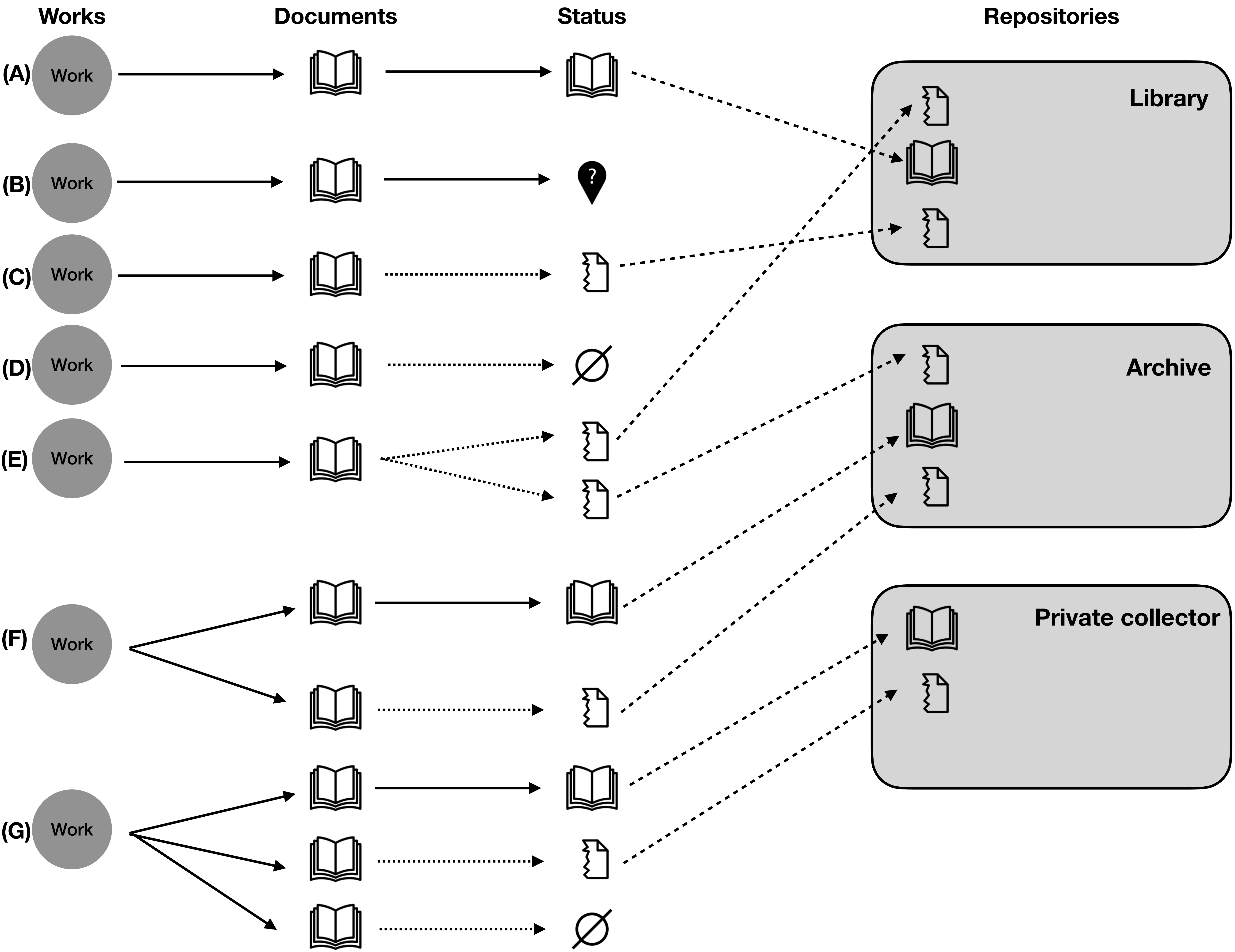
12. D. Ó Corráin, What happened to Ireland's medieval manuscripts? *Peritia*. **22–23**, 191–223 (2011–2012).
13. R. Wilson, *The Lost Literature of Medieval England* (Methuen, 1970).
14. P. Eggert, *The Work and the Reader in Literary Studies: Scholarly Editing and Book History* (Cambridge University Press, 2019).
15. H. Wijsman, *Luxury Bound. Illustrated Manuscript Production and Noble and Princely Book Ownership in the Burgundian Netherlands (1400-1550)* (Brepols, 2010).
16. J. L. Cisne, How science survived: Medieval manuscripts' "demography" and classic texts' extinction. *Science*. **307**, 1305–1307 (2005).
17. G. Declercq, Comment on "How Science Survived: Medieval Manuscripts' 'demography' and Classic Texts' Extinction." *Science*. **310**, 1618–1618 (2005).
18. N. D. Pyenson, L. Pyenson, E. Buringh, J. L. Cisne, Treating medieval manuscripts as fossils. *Science*. **309**, 698–701 (2005).
19. M. S. Cuthbert, Tipping the iceberg: Missing Italian polyphony from the age of Schism. *Musica Disciplina*. **54**, 39–74 (2009).
20. L. Egghe, G. Proot, The estimation of the number of lost multi-copy documents: A new type of informetrics theory. *Journal of Informetrics*. **1**, 257–268 (2007).
21. Materials and methods are available as supplementary materials.
22. A. Chao, Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*. **11**, 265–270 (1984).
23. A. Chao, C. H. Chiu, Estimation of species richness and shared species richness, in *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, N. Balakrishnan, Ed. (Wiley, 2012), pp. 76–111.
24. M. I. Eren, A. Chao, W.-H. Hwang, R. K. Colwell, Estimating the richness of a population when the maximum number of classes is fixed: A nonparametric solution to an archaeological problem. *PLOS ONE*. **7**, 1–11 (2012).
25. A. Chao, Y. Kubota, D. Zelený, C.-H. Chiu, C.-F. Li, B. Kusumoto, M. Yasuhara, S. Thorn, C.-L. Wei, M. J. Costello, R. K. Colwell, Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*. **35**, 292–314 (2020).
26. M. O. Hill, Diversity and evenness: A unifying notation and its consequences. *Ecology*. **54**, 427–432 (1973).
27. A. Chao, R. K. Colwell, C.-W. Lin, N. J. Gotelli, Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*. **90**, 1125–1133 (2009).
28. A. Chao, C. Ricotta, Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*. **100**, e02852 (2019).
29. I. Donohue, H. Hillebrand, J. M. Montoya, O. L. Petchey, S. L. Pimm, M. S. Fowler, K. Healy, A. L. Jackson, M. Lurgi, D. McClean, N. E. O'Connor, E. J. O'Gorman, Q. Yang, Navigating the complexity of ecological stability. *Ecology Letters*. **19**, 1172–1185 (2016).
30. R. J. Whittaker, J. M. Fernández-Palacios, *Island Biogeography. Ecology, Evolution, and Conservation* (Oxford University Press, 2006).
31. A. Acerbi, R. A. Bentley, Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior*. **35**, 228–236 (2014).
32. M. Cantor, M. Chimento, S. Q. Smeele, P. He, D. Papageorgiou, L. M. Aplin, D. R. Farine, Social network architecture and the tempo of cumulative cultural evolution. *Proceedings of the Royal Society of London, Series B: Biological Sciences* (2021), doi:10.1101/2020.12.04.411934.

33. D. Kaufman, Measuring archaeological diversity: An application of the jackknife technique. *American Antiquity*. **63**, 73–85 (1998).
34. W. W. Esty, How to estimate the original number of dies and the coverage of a sample. *The Numismatic Chronicle*. **166**, 359–364 (2006).
35. L. C. Briand, K. El Emam, B. G. Freimut, O. Laitenberger, A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Transactions on Software Engineering*. **26**, 518–540 (2000).
36. K. G. Janardan, D. J. Schaeffer, Methods for estimating the number of identifiable organic pollutants in the aquatic environment. *Water Resources Research*. **17**, 243–249 (1981).
37. E. B. Hook, R. R. Regal, Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic reviews*. **17**, 243–264 (1995).
38. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn, C. F. Weber, Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*. **75**, 7537–7541 (2009).
39. S.-P. Huang, B. S. Weir, Estimating the total number of alleles using a sample coverage method. *Genetics*. **159**, 1365–1373 (2001).
40. B. Efron, R. Thisted, Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*. **63**, 435–447 (1976).
41. K. P. Burnham, W. S. Overton, Robust estimation of population size when capture probabilities vary among animals. *Ecology*. **60**, 927–936 (1979).
42. C.-H. Chiu, Y.-T. Wang, B. A. Walther, A. Chao, An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics*. **70**, 671–682 (2014).
43. I. J. Good, The population frequencies of species and the estimation of population parameters. *Biometrika*. **40**, 237–264 (1953).
44. A. Chao, R. K. Colwell, Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *Statistics and Operation Research Transactions*. **41**, 3–54 (2017).
45. A. Chao, Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*. **43**, 783–791 (1987).
46. A. Chao, L. Jost, Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*. **6**, 873–882 (2015).
47. R. K. Colwell, J. A. Coddington, Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. **345**, 101–118 (1994).
48. D. Böhning, Chao's lower bound estimator and the size of the Pleiades. *Environmental and Ecological Statistics*. **27**, 171–173 (2020).
49. M. Kestemont, F. Karsdorp, Het Atlantis van de Middelnederlandse ridderepiek. Een schatting van het tekstverlies met methodes uit de ecodiversiteit. *Spiegel der Letteren*. **61**, 271–290 (2019).
50. M. Kestemont, F. Karsdorp, Estimating the loss of medieval literature with an unseen species model from ecodiversity, in *Proceedings of the workshop on computational humanities research (CHR 2020), Amsterdam, the Netherlands, November 18-20, 2020*, F. Karsdorp, B.

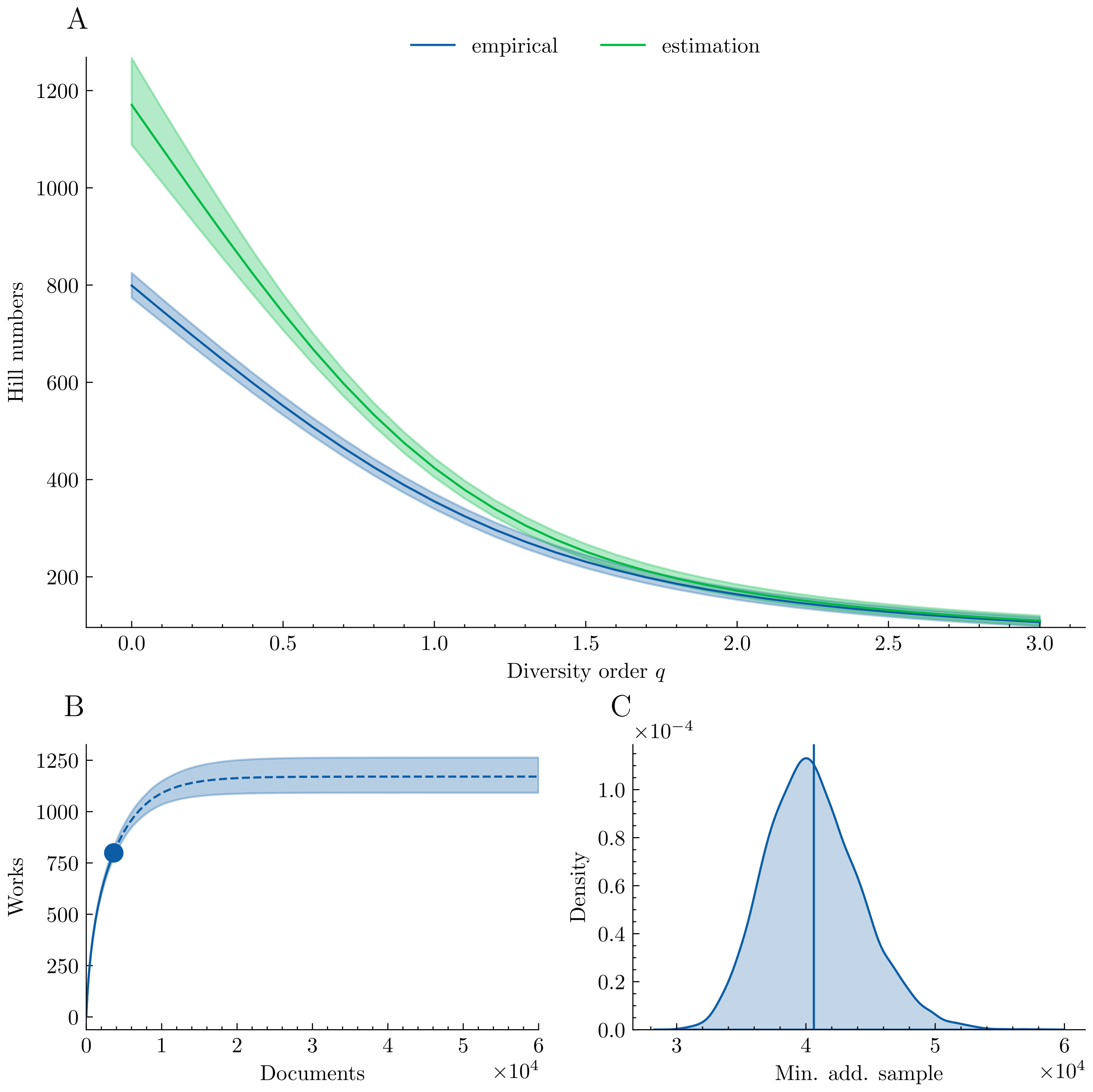
McGillivray, A. Nerghes, M. Wevers, Eds. (CEUR-WS.org, 2020), vol. 2723 of *CEUR Workshop Proceedings*, pp. 44–55.
51. Q. L. Burrell, Some comments on "The estimation of lost multi-copy documents: A new type of informetrics theory" by Egghe and Proot. *Journal of Informetrics*. **2**, 101–105 (2008).
52. J. Green, F. McIntyre, P. Needham, The shape of incunable survival and statistical estimation of lost editions. *The Papers of the Bibliographical Society of America*. **105**, 141–175 (2011).
53. H. Hillebrand, D. M. Bennett, M. W. Cadotte, Consequences of dominance: A review of evenness effects on local and regional ecosystem processes. *Ecology*. **89**, 1510–1520 (2008).
54. R. MacArthur, Fluctuations of animal populations and a measure of community stability. *Ecology*. **36**, 533–536 (1955).
55. N. G. Hairston, J. D. Allan, R. K. Colwell, D. J. Futuyma, J. Howell, M. D. Lubin, J. Mathias, J. H. Vandermeer, The relationship between species diversity and stability: An experimental approach with protozoa and bacteria. *Ecology*. **49**, 1091–1101 (1968).
56. S. L. Pimm, The complexity and stability of ecosystems. *Nature*. **307**, 321–326 (1984).
57. D. U. Hooper, F. Chapin III, J. J. Ewel, A. Hector, P. Inchausti, S. Lavorel, J. H. Lawton, D. Lodge, M. Loreau, S. Naeem, others, Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological monographs*. **75**, 3–35 (2005).
58. D. Tilman, P. B. Reich, J. M. Knops, Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*. **441**, 629–632 (2006).
59. E. C. Pielou, The measurement of diversity in different types of biological collections. *Journal of theoretical biology*. **13**, 131–144 (1966).
60. E. R. Crema, A. Kandler, S. Shennan, Revealing patterns of cultural transmission from frequency data: Equilibrium and non-equilibrium assumptions. *Scientific Reports*. **6** (2016), doi:10.1038/srep39122.
61. L. S. Premo, Effective population size and the effects of demography on cultural diversity and technological complexity. *American Antiquity*. **81**, 605–622 (2016).
62. M. Derex, R. Boyd, Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*. **113**, 2982–2987 (2016).
63. W. J. Ewens, *Mathematical Population Genetics 1: Theoretical Introduction* (Springer Science & Business Media, 2012), vol. 27.
64. A. Kandler, B. Wilder, L. Fortunato, Inferring individual-level processes from population-level patterns in cultural evolution. *Royal Society Open Science*. **4**, 170949 (2017).
65. E. Buringh, The Population of European Cities from 700 to 2000, *Research Data Journal for the Humanities and Social Sciences*. 2021, DOI:https://doi.org/10.1163/24523666-06010003.
66. M. Clanchy, *From Memory to Written Record: England 1066–1307 (3rd Ed.)* (Wiley, 2012).
67. P. Zumthor, Ed., *Essai de poétique médiévale* (Seuil, 1972).
68. E. R. Curtius, Ed., *Europäische Literatur und Lateinisches Mittelalter* (Francke Verlag, 1948).
69. E. Auerbach, Ed., *Mimesis: Dargestellte Wirklichkeit in der abendländischen Literatur* (Francke Verlag, 1946).
70. L. B. Mortensen, European Literature and Book History in the Middle Ages, c. 600-c. 1450, in *Oxford Research Encyclopedias* (Oxford University Press, 2018).
71. T. Tanselle, *A Rationale of Textual Criticism* (University of Pennsylvania Press, 1989).

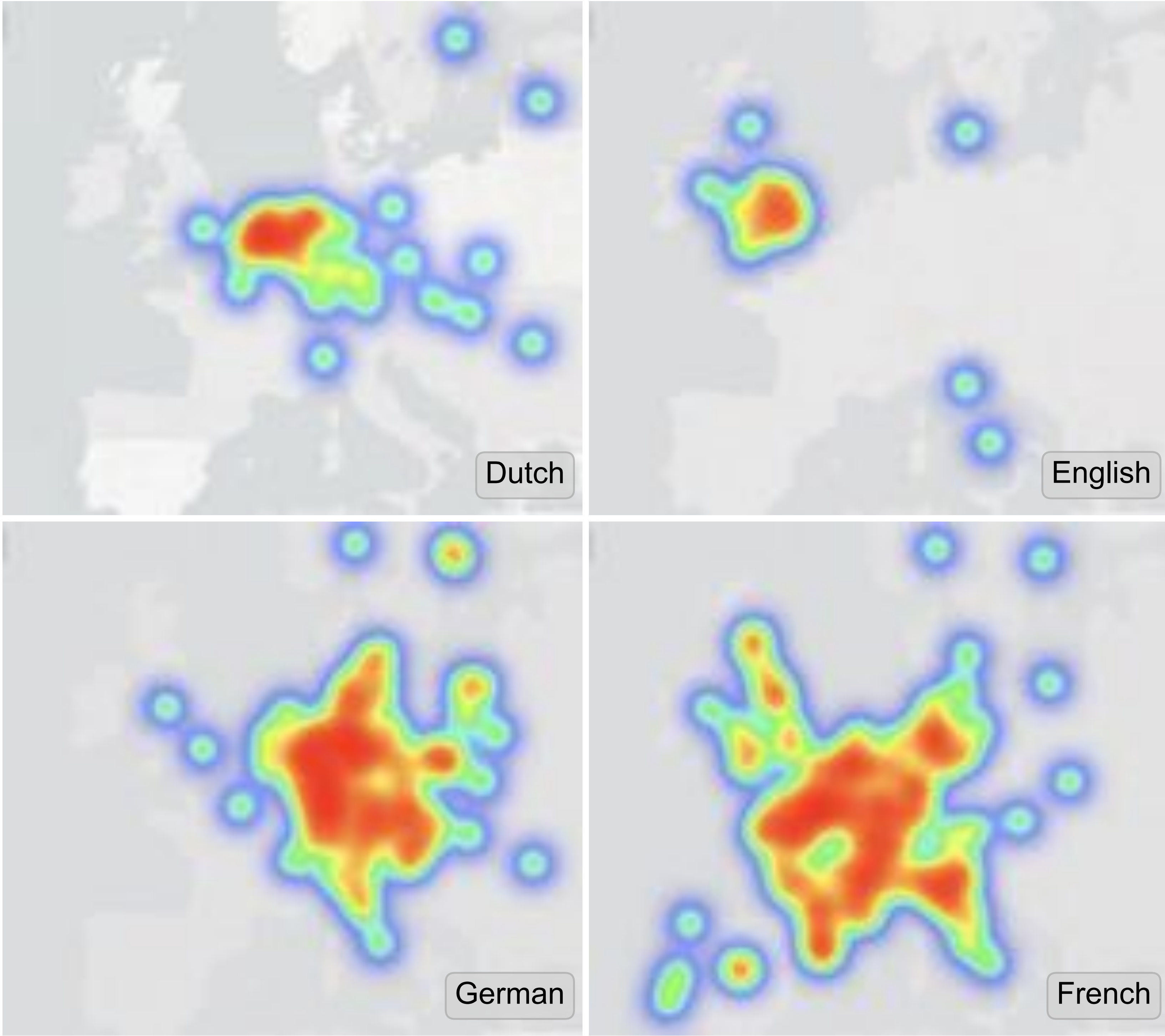
72. P. Shillingsburg, *Textuality and Knowledge: Essays* (University of Pennsylvania Press, 2017).
73. A. C. Barbrook, C. J. Howe, N. Blake, P. Robinson, The phylogeny of The Canterbury Tales. *Nature*. **394**, 839 (1998).
74. P. Roelli, Ed., *Handbook of Stemmatology: History, Methodology, Digital Approaches* (De Gruyter, 2020).
75. D. Sawyer, Verse-Craft, editing, and the work: Shadows of Orfeo. *The Review of English Studies* (2021), doi:10.1093/res/hgab058.
76. F. Van Oostrom, *Stemmen op schrift. Geschiedenis van de Nederlandse literatuur van het begin tot 1300* (Prometheus, 2006).
77. H. Kienhorst, *De handschriften van de Middelnederlandse ridderepiek. Een codicologische beschrijving* (Sub Rosa, 1985).
78. J. W. Klein, "Het getal zijner jaren is onnaspeurlijk." Een herijking van de dateringen van de handschriften en fragmenten met Middelnederlandse ridderepiek. *Tijdschrift voor Nederlandse Taal- en Letterkunde*. **111**, 1–23 (1995).
79. B. Caers, Een *buchelin inn flemische*. Over ontstaan en verspreiding van de ridderepiek in de Nederlanden (ca. 1150-1450). *Tijdschrift voor Nederlandse Taal- en Letterkunde*. **127**, 223–251 (2011).
80. B. Caers, M. Kestemont, De datering van de Middelnederlandse ridderepiek. *Verslagen en mededelingen van de Koninklijke Academie voor Nederlandse taal- en letterkunde*. **121**, 1–59 (2011).
81. M. Zink, *Littérature française du Moyen Age* (Presses universitaires de France, 1992).
82. B. Bastert, *Helden als Heilige: Chanson de geste-Rezeption im deutschsprachigen Raum* (Francke, 2010).
83. E. Lienert, *Deutsche Antikenromane des Mittelalters* (Erich Schmidt Verlag, 2001).
84. W. Achnitz, *Deutschsprachige Artusdichtung des Mittelalters: eine Einführung* (De Gruyter, 2012).
85. H. Brunner, *Geschichte der deutschen Literatur im Überblick* (Reclam, 1997).
86. R. H. Bremmer, Old English heroic literature, in *Readings in Medieval Literature. Interpreting Old and Middle English Literature*, D. F. Johnson, E. Treharne, Eds. (Oxford University Press, 2005), pp. 75–90.
87. J. Wogan-Browne, Ed., *Language and Culture in Medieval Britain: The French of England, c.1100–c.1500* (Woodbridge, 2009).
88. S. Crane, *Insular romance: Politics, Faith, and Culture in Anglo-Norman and Middle English Literature* (University of California Press, 1986).
89. R. J. Dean, *Anglo-Norman Literature: A Guide to Texts and Manuscripts* (Anglo-Norman Text Society, 1999).
90. T. H. Tulinius, *The Matter of the North: The Rise of Literary Fiction in Thirteenth-century Iceland* (Odense University Press, 2002).
91. M. Kalinke, P. Mitchell, Eds., *Bibliography of Old Norse-Icelandic romances* (Cornell University Press, 1985).
92. P. E. Ólason, L. H. Blöndal, G. M. Helgason, Ö. Helgason, Eds., *Skrá um handritasöfn Landsbókasafnsins* (Landsbókasafn Íslands, 1918-1996).
93. K. Kålund, Ed., *Katalog over den Arnamagnæanske håndskriftsamling* (Gyldendalske boghandel, 1889-1894).
94. F. R. Sigmundsson, *Rímnatal* (Rímnafélagið, 1966).

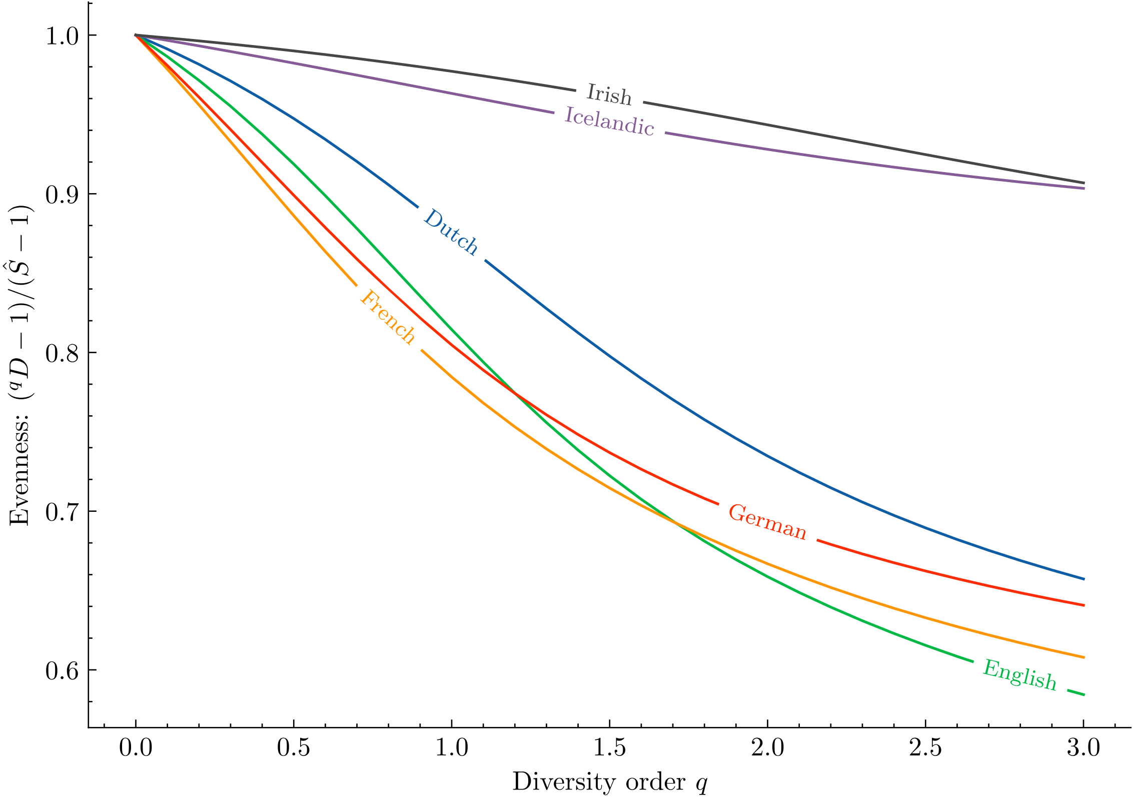
95. B. K. Thórólfsson, *Rímur fyrir 1600. Safn fræðafjelagsins um Ísland og Íslendinga* (S.L. Möller, 1934).
96. F. Jónsson, Ed., *Rímnasafn: Samling af de ældste Islandske Rimer* (Samfund til udgivelse af gammel nordisk litteratur, 1905–1922).
97. D. Ó Corráin, *Clavis litterarum Hibernesium: Medieval Irish Books & Texts (c. 400–1600)* (Brepols, 2017).
98. T. Burrows, C. Johnston, Eds., *Collecting the Past: British Collectors and their Collections from the 18th to the 20th Centuries* (Routledge, 2019).



Works
Documents
Status
Repositories
(A)
(B)
(C)
(D)
(E)
(F)
(G)
Work
Work
Work
Work
Work
Work
Work
?
Library
Archive
Private collector


A
empirical
estimation
Hill numbers
1200
1000
800
600
400
200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Diversity order q
B
Works
1250
1000
750
500
250
0
0
1
2
3
4
5
6
Documents
$\times 10^4$
C
$\times 10^{-4}$
Density
1.0
0.8
0.6
0.4
0.2
0.0
3
4
5
6
Min. add. sample
$\times 10^4$


Dutch
English
German
French



Irish
Icelandic
Dutch
French
German
English
1.0
0.9
0.8
0.7
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Evenness: $({}^qD - 1)/(\hat{S} - 1)$
Diversity order q