Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity

Mike Kestemont[a], Folgert Karsdorp[b]

[a] *Department of Literature, University of Antwerp, Antwerp, Belgium*
[b] *Royal Netherlands Academy of Arts and Sciences, Meertens Institute, Amsterdam, The Netherlands*

Abstract

The century-long loss of documents is one of the major impediments to the study of historic literature. Here we focus on Middle Dutch chivalric epics (ca. 1200-1450), a genre for which little archival records exist that shed light on the survival rates of works and documents. We cast the quantitative estimation of these survival rates as a variant of the unseen species problem from ecodiversity. We apply an established non-parametric method (Chao1) and compare it to a number of common alternatives on simulated data. Finally, we discuss the implications of our results for conventional philology: our numbers suggest that the losses sustained on the level of works may be more dramatic than previously imagined, whereas those at the document-level align surprisingly well with existing estimates in book history, although these were based on completely different data sources.

Keywords

medieval literature, book history, unknown species problem, Middle Dutch, ecodiversity

1. Introduction: the survival of premodern literature

The century-long loss of material artifacts is one of the major impediments to the study of the history of human culture. Across various domains in the humanities, scholars must base their study on incomplete archival collections that offer but a tiny fraction of the wealth of historical specimens that originally existed. In this contribution, we focus on the domain of literature from the High Medieval period in Western Europe, which has sustained significant losses in the past centuries. Previous work has argued that unseen species models from ecodiversity can be used to estimate the number of works (multi-copy documents) that have been lost to us [e.g., 16, 26]. Although these models have already yielded interesting insights for early modern printed works, they have hardly been applied to premodern handwritten literature so far [exceptions include 13, 26]. Here, we apply Chao1, a non-parametric estimator of asymptotic species richness, to a representative corpus of Middle Dutch romances and quantitatively evaluate its performance on simulated datasets. A novelty of this contribution is that we do not only estimate the proportion of lost works in this dataset, but also the number of lost documents, through an extension of Chao1, which aims to gauge the number of additional samples that would be minimally required to reach the asymptote of the species accumulation curve, estimated by the original model.

In traditional philology, a theoretical distinction is typically drawn between the abstract notion of a "work" and the physical "documents" (witnesses, carriers) in which the work is

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

mike.kestemont@uantwerpen.be (M. Kestemont); folgert.karsdorp@meertens.knaw.nl (F. Karsdorp)
0000-0003-3590-693X (M. Kestemont); 0000-0002-5958-0551 (F. Karsdorp)

attested in some version [36]. Throughout the Middle Ages (ca. 500-1500 AD), handwritten media, such as manuscripts or scrolls, were the primary physical medium for the sustainable exchange of literary texts [28, 2]. Before the advent of printing, all witnesses of a text were hand-copied from pre-existing exemplars, a practice that yielded textual traditions in which intricate interdependencies exist between copies. The document tree resulting from this process is known as the *stemma codicum* and such trees are nowadays studied in the field of phylogenetics [1].

Medieval text traditions, however, rarely survive in full, as many documents have now been lost, due to a variety of historical reasons, including natural or infrastructural disasters (e.g. library fires, such as the famous example of Alexandria) but also the wilful destruction by humans, such as the controlled disposition of duplicates by heritage institutions or collectors [2]. Moreover, many sources have only survived fragmentarily and often the severely damaged remnants of the same book are nowadays even scattered across various locations.[1] This is related to the fact that, in the premodern period, book binders regularly recycled parchment codices into "maculature" that was used, for instance, to strengthen the spines of newer books, which eventually ended up in different locales [19].

We can assume that a large fraction of premodern documents, if not an absolute majority, is nowadays unknown to us [32], either because the documents no longer exist, or because they have not been recovered yet (e.g., due to cataloguing initiatives that are lagging behind) [22]. Consequently, a great deal of *works* are also unknown to us, in the obvious case where all the documents representing a work are currently unknown [40, 24]. These assumptions are not only justified by the many references in historic sources to works that we no longer know, but also by the constant stream of new material findings nowadays – which has clearly been intensified in recent decades by the emergence of the internet and social media [21]. Understanding the literary preferences of the past, and explaining historical shifts therein, is one of the core tasks of cultural studies and a prerequisite for producing valid literary histories. Nevertheless, it is clear that the situation of partial observability, outlined above, severely compromises this task: the available data only constitute a very limited sample of an original population of literature that was much larger and more diverse. In statistical terms, our present-day perspective is by necessity biased towards the materials that actually survived. Understandably, scholars invariably agree that it is vital to correct these biased preconceptions and account for the materials which are are no longer known to us [40, 24, 2].

Methodologically, it is important to separate the loss of documents, from the loss of works which it entails. As to the second matter, the loss of works, there has been very little empirical work in the field of medieval studies, beyond the descriptive analysis of historic references to lost works [40, 24]. There has been some empirical research into the first matter. Book historical studies (such as [2]) have mainly studied the survival rate of documents on the basis of the limited set of medieval collections, of which the composition is exactly known at specific points in time. This allows one to quantify the gradual, diachronic loss of documents from these collections. While these estimates are currently among the best we have, it is clear that it can be hard to extrapolate these numbers to other regions, languages or collection environments (e.g. monastic vs. lay book possession), so that alternative approaches to complement this methodology would be a valuable addition to the field. Finally, it is worth mentioning the

[1] A dramatic example is the Beauvais Missal, of which the dismembered folios are currently being pieced together again in a virtual reconstruction. Updates on the Broken Books project by Lisa Fagin Davis can be followed here: https://web.archive.org/save/https://brokenbooks2.omeka.net/.

polemic that ensued the 2005 high-profile publication by John L. Cisne in *Science* [11]. This paper used methods from geology and population biology to estimate the rate of manuscript loss for a set of early medieval text traditions, but has almost instantly been met with severe, yet well-founded criticism from a number of well-placed medievalists [15, 35].

2. Related work: bibliometry and the unknown species problem

In a pioneering contribution, Egghe & Proot [16] have proposed a probabilistic model that attempts to estimate the level of loss for historic works (multi-copy documents), based on the frequency with which retrieved copies of such works survive. Their case study was based on bibliometric data, drawn from a short-title catalogue of printed works from the Low Countries. Follow-up work has confirmed the practical usefulness of their approach for printed works [34, 23, 22, 33]. Their model can be formalized as follows:

$$\hat{f}_0 = \left(\frac{1}{1 + \frac{2f_2}{(a-1)f_1}} \right)^a \tag{1}$$

In this formula, f_1 is the number of works in a given corpus that survive in exactly one copy and f_2 the number of works that survive in exactly two copies; a is the number of copies that were produced of each work, which is the so-called "run" for printed works (which they set to 500 copies). These coefficients are then used to estimate $\hat{f}_0$, or the proportion of lost works from the total, original population of works. In a sagacious response to Egghe & Proot [16], Burrell [4] has noted that their task could be considered as a variant of a much older problem, namely the "unseen species problem". This problem is studied in various fields, ranging from ecology to genetics, where scholars have to estimate aspects of species diversity (e.g. biota richness) in a specific assemblage on the basis of highly incomplete samples of the full population [14]. This task has a rich tradition in biostatistics, reaching back to the 1940s, with the work of Alexander Steven Corbet, who had been trapping and inventorizing new butterflies species in British Malaya for two years [31]. In collaboration with the statistician R.A. Fisher [18], he formulated a model to estimate the number of new species he would discover, if he were to continue his trapping efforts for another two years.

Nowadays there exists a variety of statistical approaches to the unseen species problem that can be borrowed from ecodiversity, an interdisciplinary domain where researchers study, amongst other things, the biota richness in ecosystems. Monitoring the number of unique species, for example, is a key task for various environmental reasons, for instance, to assess the impact of natural disasters [14]. These approaches are well established [29, 7, 20] but not all of these are applicable to our kind of data. Applying the pioneering model by Egghe & Proot [16], for instance, is not without theoretical issues, because the concept of a print run is almost meaningless in this context (cf. the a coefficient in Eq. 1), even though the authors show that its effect is limited. The serial production of handwritten text carriers was extremely uncommon throughout the Middle Ages, as books were still highly customized luxury objects that were never mass-produced. It seems impossible to provide an estimate for this parameter, also because the available evidence suggests that the number of original copies per work was heavily skewed (dependent on factors such as genre, language or general prestige). It is likely that the large majority of medieval works already originally only existed in very few copies (i.e. singletons or doubletons).

To a reasonable extent, these methodological caveats are mitigated by the Chao1 estimator [6, 5], a non-parametric method that is robust (even universally valid) in the face of unknown species abundance distributions and enables the comparison of species richness across multiple assemblages [7]. Previous, exploratory work in literary studies [26] has shown that this estimator produces interesting results for handwritten sources. This method is especially attractive for highly diverse, log-normally distributed assemblages, typical of human cultural production, where many species are infrequent and thus hard to detect. In such cases, it is futile to try and offer a precise point estimate; Chao1 therefore rather offers an accurate *lower bound* of the number of undetected species in a sample. The estimator is given by [8]:

$$\hat{f}_0 = \begin{cases} \dfrac{(n-1)}{n}\dfrac{f_1^2}{(2f_2)} & iff_2 > 0; \\ \dfrac{(n-1)}{n}\dfrac{f_1(f_1-1)}{2} & iff_2 = 0 \end{cases} \tag{2}$$

Here, f_1 is the number of species sighted exactly once in the sample (singletons), f_2 the number of species that were sighted twice (doubletons), and n the observed, total sample size (cf. Eq. 1). Finally, $\hat{f}_0$ is the estimated lower bound for the number of species that do exist in the assemblage, but which were sighted zero times, i.e. the number of undetected species. To obtain a confidence interval, a simple bootstrap procedure can be applied, in which the available data is iteratively resampled [8].

An attractive feature of this estimator is that it can be naturally extended to estimate the number of lost documents (instead of the number of lost works) [10]. Field workers tasked with biodiversity sampling often do not observe a substantial fraction of the biota that live in a certain assemblage. While Chao1 can estimate how many of this low-abundance species have (minimally) gone undetected, it does not tell us how much additional effort would be required to observe these, i.e. how many additional m individuals would have to be sampled to observe all of the biota *at least once*. Put informally, with respect to the species accumulation curve (cf. Fig. 2), we would like to find out in which area the asymptote starts to kick in.

Using the same abundance data as above, this extension of Chao1 tries to estimate at which point every species would have been observed at least as a singleton. The singletons in the enlarged sample of size $m + n$ (where n is still the number of previously observed individuals in the sample) would fall apart in two distinct categories [10]: (1) singletons from the original sample, for which no additional individuals are detected by the enlarged sample, and (2) previously undetected species for which exactly one individual is observed during the additional sampling. The estimator aims to calculate the proportion between (1) and (2) to determine m on the basis of two functions. The first function, $h(x) = 2f_1(1 + x)$, is a linear transformation of x, whereas the second function, $v(x) = exp[x(2f_2/f_1)]$, is an exponentially increasing function; v is bound to intersect h at a certain $x* > 0$. The number of additional m individuals that are theoretically required to observe the full richness of a population is given by: $m = nx*$. Here too, a bootstrapping procedure can be used to estimate a confidence interval.

Regarding historic literature, the analogy in applying this method is straightforward: how many additional documents would have to be rediscovered in the future to observe all works at least once? While this estimate has very useful, practical implications for philologists scanning archives for new fragments, the resulting number, $m + n$, also has theoretical relevance, since it would be reasonably close to the actual size of the original population of documents. Thus, $m + n$ would allow us to estimate the historic loss of documents, based on a type of data that is complementary to (and even completely independent from) the archival library records mentioned above. Because of the log-normal distribution of literary works over documents, we expect that most works were of an extremely low historic abundance, i.e. they were already originally produced in very low numbers of copies. We can therefore assume that the majority of works that are currently unknown will in the future only be detected in a low number of documents. Once we would have observed all works, we can therefore expect that we would also have observed most documents. Thus, while the outcome of this method should not be treated as a precise point estimate – like Chao1, it too estimates the *minimal* sampling effort required – we argue that it offers a useful approximation of the historical loss of documents. Nevertheless, we should emphasize that this method likely yields an underestimation of the original document richness and it would not account for specific aspects of the historic document mass, for instance, in cases where presently unknown works actually survive in more than one (so far undetected) documents.

3. Estimating the loss of Middle Dutch chivalric epics

We have collected the surviving works and documents from the genre of Middle Dutch chivalric epics (*ridderepiek*) as abundance data, where we record in how many documents a particular work has been "sighted". This data is mainly drawn from Kienhorst's acclaimed repertory [27] but we have updated this information with newer, and even very recent findings (situation as of 10 July 2020).[2] The main bibliographic information can be gleaned from Table 1, showing, in the last row, how the 75 presently known works are distributed over the 167 documents (=n) that have been retrieved. 45 works are attested as *unica* in only a single source (=$f1$); 13 works are doubletons (=$f2$).

works	documents
45	1
13	2
6	3
2	4
4	5
1	6
1	7
2	10
1	17
75	167

Table 1: Abundance data for the Middle Dutch chivalric epics. Last row are the total counts.

3.1. Loss of works

We can plug these numbers into the equations presented above and arrive at the estimates presented in Table 2. Here, we additionally give the estimate of the so-called Jackknife procedure (following the reference implementation in [38]), a historic alternative to the more recent estimators that generally aim to reduce the bias in estimators [3]. Importantly, this approach lacks theoretical justification [12] but offers a surprisingly solid baseline in many practical applications [7, 29]. We mention in passing that such techniques have already been applied in domains that border on the Humanities, such as archaeology [25, 17]. We also present the confidence intervals (CI) obtained from the bootstrap procedure: these are fairly wide but show considerable overlap, thus stressing the relative agreement between the three estimators. The distribution of the bootstrap values is shown in the rainplots (Fig. 1), except for the Jackknife (for which the CI is calculated analytically). We observe that Chao1 gives the most conservative estimate for the loss of works, which was to be expected, given the fact that it estimates a lower bound for the loss. The Jackknife and EP procedure both estimate a higher loss rate (yet both in the same range).

[2] Data and code supporting this paper have been publicly archived: https://doi.org/10.5281/zenodo.4030681.

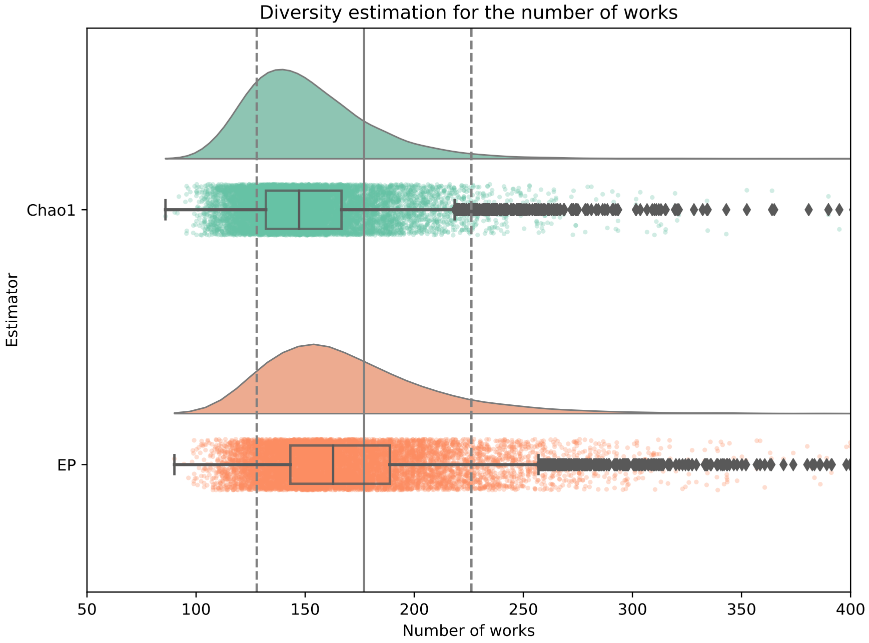

Figure 1: Distribution of bootstrap estimations for Chao1 and EP on the Middle Dutch data. The Jackknife estimate (with its CI) is added with vertical lines.

Crucial for the discussion below, is that all three estimators for the loss of works suggest that only half (and potentially even less) of the original works that once existed are currently known to us.

3.2. Loss of documents

Method	Estimate	CI
Chao1	152.42	110.11 - 222.98
Jackknife	177.00	127.81 - 226.19
EP	170.71	116.77 - 268.49
Minsample	2047.77	1064.19 - 4006.42

Table 2: Diversity estimates for the Middle Dutch chivalric epics (with CI). Last row is the result for minimum additional sampling.

The final row in Table 1 gives the estimate (with CI) for the loss of documents. While we should account for an extremely wide CI in this case, the number suggest a survival rate of ≈8.15%, i.e. of an original population of 2047 documents, only 167 have survived. We offer a final and joint visualization of the results in Fig. 2. This plot shows what is known as a “species accumulation curve” [9]. The blue line plots the number of retrieved works as an (asymptotic) function of the number of documents recovered in this assemblage. The full line indicates the situation for the observed sample, whereas the dashed part concerns the hypothetical increase, in the case where more “sightings” would occur in the future. The grey distribution shows the bootstrap values resulting from the minimum sampling effort estimator, broadly indicating the region where we expect the curve to hit the asymptote.

The green distribution in Fig. 2 requires additional explanation. The available models for estimating the population size (as opposed to species diversity) of an under-sampled assemblage typically assume that we have capture-recapture information available, instead of mere

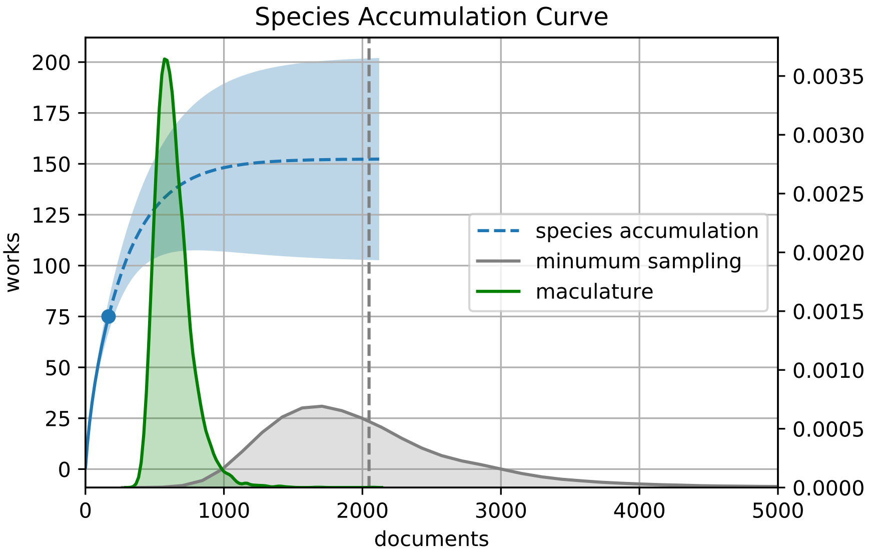

Figure 2: Species accumulation curve (blue), including bootstrap distribution for minimum additional document sampling (in grey; the dashed vertical grey line indicates the non-bootstrapped estimate) and for the maculature diversity (in green).

abundance data [5]. We cannot extract such information from our data – because a work-document pair can in principle only be “sighted” once and after that it is not released again “into the wild”. Nevertheless, in the case of manuscripts that have been recycled into maculature, the remnants of the same document have often reappeared in different locations – an extreme example is the *Roman der Lorreinen*-codex of which 9 fragments resurfaced, scattered across 7 different libraries [27]. We can apply Chao1 to the documents in our corpus that survive fragmentarily and represent them as abundance data, on the basis of the number of fragments that resurfaced of them. This yields an assemblage of 141 documents surviving in 181 fragments, with $f1 = 118$ and $f2 = 14$. The application of Chao1 yields the following estimate: 635.54 CI(449.85 - 947.25) (cf. the green area in Fig. 2). Note that this number does not estimate the total number of documents that once existed, but rather the size of the subset of manuscripts that were recycled into maculature. In combination with the other estimate, our analyses suggest that ≈31% of the original population of documents with chivalric Middle Dutch epics was recycled into maculature.

4. Simulations

In this section, we compare the performance of the three estimators for species diversity using simulated data. In the aforementioned seminal paper [18], Fisher proposed to model the abundance of species in an assemblage as $S_n = \alpha x^n/n$, where S_n is the number of species with an abundance of n, x a positive constant ($0 < x < 1$) which generally approaches 1 and α is the number of singleton species in the assemblage. This logseries is still in wide use and and can be used to define a discrete probability distribution, parameterized by two values: (i) the number of singleton species in the population and (ii) the maximum abundance for a single species (to put a practical cap on the distribution).

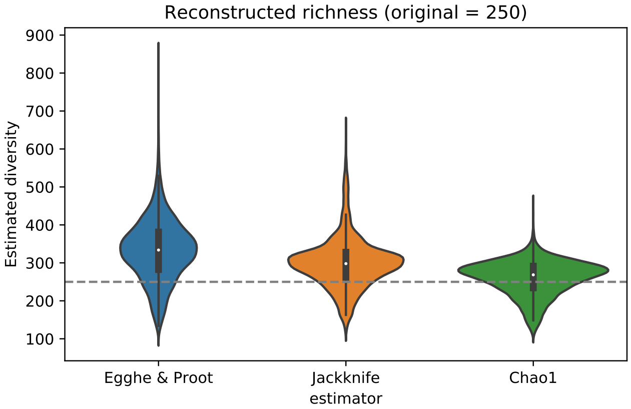

Figure 3: Results for the three estimators (with $\alpha = 50$ for the Egghe & Proot method) for artificial assemblages of 250 works (see dashed vertical grey line) that were stochastically downsampled.

In an iterative process, we have generated assemblages from a logseries distribution for 250 works, for a fixed $f1 = 75$ and $x = .99$. Next, we mimicked a distribution of these works over a variable number of documents (in a linear range [500, 2500]). We then modelled historic document loss as a fully stochastic process, in which documents are randomly dropped at a certain loss rate (in the linear range [0.05-0.95]). We repeated each experiment 50 times with different random seeds. We can then assess the performance of each estimator with respect to the ground truth of 250 works. The violin plots in Fig. 3 show that Chao1 is the most conservative evaluator that generally realizes the smallest deviation from the ground truth (cf. dashed grey line). Fig. 4 plots the absolute error per estimator as a function of the varying loss rate. Here, we see that Chao1 is most robust estimator throughout, except for extremely small document keep rates (< 0.1).

5. Discussion

In his acclaimed 2006 history of Middle Dutch literature, Van Oostrom estimated that the corpus of Middle Dutch chivalric epics must originally have comprised “at least 100 texts” [37]. All the estimators considered here agree that this in all likelihood too low an estimate: it seems likely that at the very least 152 texts once existed, and potentially even more, of which only 75 (≈49%) survive now, providing an even firmer basis to the claims from a previous study [26]. Chao1 proved a more reliable estimator in our simulations than the other two methods studied here, which tended to overshoot and thus overestimate the loss of works. Middle Dutch studies might have overestimated the representativeness of the surviving corpus, and future studies should attempt to account for this bias.

Although there are few previous estimates regarding the loss of works, we are on more solid grounds regarding the loss of documents. In book history, scholars have studied the loss rates for medieval documents, based on data for the sparse set of manuscript collections of which the historic composition is known, so that they can be compared to the books from these collections that are still extant today [2]. Such studies have estimated a cumulative survival rate of 7% for the sort of non-illustrated manuscripts in which Middle Dutch romances typically were copied [39, 37, 30]. While we should present these results with extreme caution for now, it

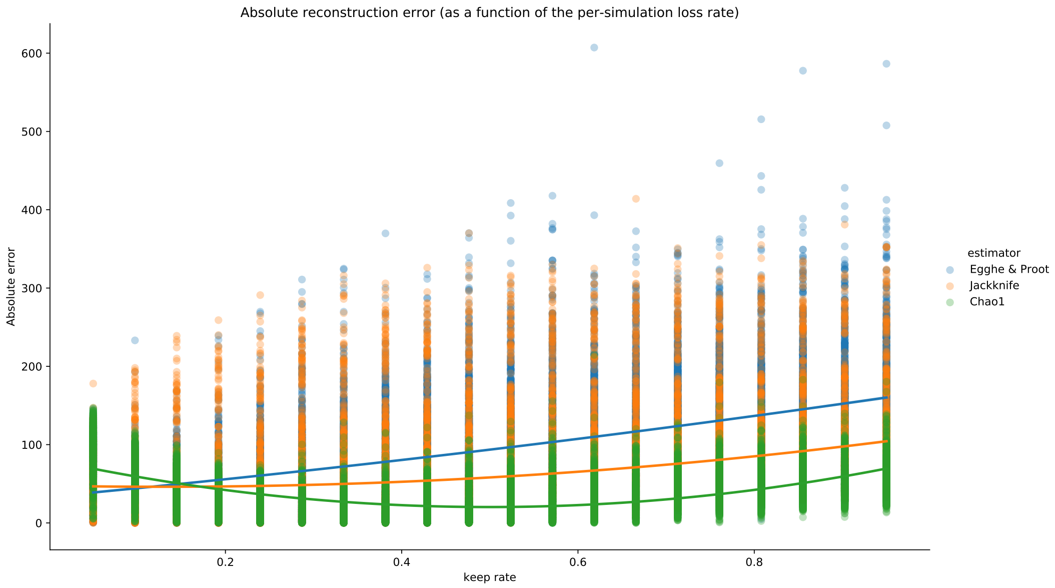

Figure 4: The absolute error for each estimator in each simulation, as a function of the loss rates considered (given a ground truth of 250), with a cubic fit per method.

is remarkable that our analysis suggest an estimate that is in a surprisingly similar range, i.e. ≈8.15% (167/2047 documents), although with a very wide CI (1064-4006). This approach might nevertheless present an exciting new research avenue that could complement the existing insights on the basis of a fully independent kind of evidence than the data used so far. Finally, our analyses suggest that of the original population of documents with chivalric Middle Dutch epics, ≈31% was recycled into maculature (i.e. 635/2047). While more research is needed to support this claim, it is the very first time to the best of our knowledge that this proportion has been estimated in a quantitative manner. This proportion is surprisingly high, which is maybe good news for the philologist, who is after all more likely to discover fragmentary sources than intact sources.

A number of issues remain with the application of these methods that require further attention. Problematic, for instance, is our assumption that document loss has been a fully stochastic process (which is the way in which we naively simulate this phenomenon here). Although there certainly are random aspects to this process, we know from traditional book history that some codices were less likely to be lost: texts in convolutes had higher survival chances, for instance, and the same has been hypothesized for higher-end (e.g. illustrated) manuscripts [39]. Future research should develop more principled, perhaps agent-based, models to simulate document loss than the fully stochastic approach adopted here. Finally, it would be interesting to extend this approach to a wider geographic and linguistic range, since these methods allow for an interesting cross-cultural comparison regarding the survival of medieval literature. This geographic variation will be a central component of our future work.

5.1. Acknowledgments

The authors would like to thank Elisabeth de Bruijn and Remco Sleiderink (University of Antwerp, BE) for the stimulating discussions, in particular about maculature. Additionally, we would like to acknowledge the helpful bibliographic input of the participants at the *Dark Archives 20/20 conference* (https://web.archive.org/save/https://aevum.space/darkarchives) where this work was previously presented.

References

[1] A. C. Barbrook et al. "The phylogeny of The Canterbury Tales". In: *Nature* 394.6696 (1998), p. 839.

[2] E. Buringh. *Medieval Manuscript Production in the Latin West, Explorations with a Global Database*. Brill, 2011.

[3] K. Burnham and W. Overton. "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals". In: *Ecology* 60.5 (1979), pp. 927–936.

[4] Q. Burrell. "Some comments on "The estimation of lost multi-copy documents: A new type of informetrics theory" by Egghe and Proot". In: *Journal of Informetrics* 2.1 (2008), pp. 101–105. ISSN: 1751-1577.

[5] A. Chao. "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability". In: *Biometrics* 43.4 (1987), pp. 783–791.

[6] A. Chao. "Nonparametric Estimation of the Number of Classes in a Population". In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 265–270.

[7] A. Chao and C.-H. Chiu. "Species Richness: Estimation and Comparison". In: Aug. 2016, pp. 1–26. ISBN: 9781118445112. DOI: 10.1002/9781118445112.stat03432.pub2.

[8] A. Chao and L. Jost. "Estimating diversity and entropy profiles via discovery rates of new species". In: *Methods in Ecology and Evolution* 6.8 (2015), pp. 873–882.

[9] A. Chao, Y. T. Wang, and L. Jost. "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species". In: *Methods in Ecology and Evolution* 4.11 (2013), pp. 1091–1100.

[10] A. Chao et al. "Sufficient sampling for asymptotic minimum species richness estimators". In: *Ecology* 90.4 (2009), pp. 1125–1133.

[11] J. L. Cisne. "How Science Survived: Medieval Manuscripts' "Demography" and Classic Texts' Extinction". In: *Science* 307.5713 (2005), pp. 1305–1307.

[12] R. M. Cormack. "Log-Linear Models for Capture-Recapture". In: *Biometrics* 45.2 (1989), pp. 395–413.

[13] M. S. Cuthbert. "Tipping the Iceberg: Missing Italian Polyphony from the Age of Schism". In: *Musica Disciplina* 54 (2009), pp. 39–74.

[14] A. Daly, J. Baetens, and B. De Baets. "Ecological Diversity: Measuring the Unmeasurable". In: *Mathematics* 6.7 (July 2018), p. 119.

[15] G. Declercq. "Comment on "How Science Survived: Medieval Manuscripts' "Demography" and Classic Texts' Extinction"". In: *Science* 310.5754 (2005), pp. 1618–1618.

[16] L. Egghe and G. Proot. “The estimation of the number of lost multi-copy documents: A new type of informetrics theory”. In: *Journal of Informetrics* 1.4 (2007), pp. 257–268. ISSN: 1751-1577.

[17] M. Eren et al. “Estimating the Richness of a Population When the Maximum Number of Classes Is Fixed: A Nonparametric Solution to an Archaeological Problem”. In: *PLOS ONE* 7.5 (May 2012), pp. 1–11.

[18] R. Fisher, A. S. Corbet, and C. Williams. “The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population”. In: *The Journal of Animal Ecology* 12.1 (1943), pp. 42–58.

[19] D. Geirnaert. ““Membra disiecta”: banden met het versneden verleden”. In: *Medioneerlandistiek. Een inleiding tot de Middelnederlandse letterkunde*. Ed. by R. Jansen-Sieben, J. Janssens, and F. Willaert. Verloren, 2000, pp. 85–101.

[20] N. J. Gotelli and A. Chao. “Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data”. In: *Encyclopedia of Biodiversity (Second Edition)*. Ed. by S. A. Levin. Second Edition. Waltham: Academic Press, 2013, pp. 195–211. ISBN: 978-0-12-384720-1.

[21] J. M. Green. “Digital manuscripts as sites of touch: using social media for ‘hands-on’ engagement with medieval manuscript materiality”. In: *Archive Journal* 6 (Sept. 2018).

[22] J. Green and F. McIntyre. “Lost Incunable Editions: Closing in on an Estimate”. In: *Lost Books. Reconstructing the Print World of Pre-Industrial Europe*. Ed. by F. Bruni and A. Pettegree. Brill, 2016, pp. 55–72.

[23] J. Green, F. McIntyre, and P. Needham. “The Shape of Incunable Survival and Statistical Estimation of Lost Editions”. In: *The Papers of the Bibliographical Society of America* 105.2 (2011), pp. 141–175.

[24] T. Haye. *Verlorenes Mittelalter: Ursachen und Muster der Nichtüberlieferung mittellateinischer Literatur*. Brill, 2016.

[25] D. Kaufman. “Measuring Archaeological Diversity: An Application of the Jackknife Technique”. In: *American Antiquity* 63.1 (1998), pp. 73–85.

[26] M. Kestemont and F. Karsdorp. “Het Atlantis van de Middelnederlandse ridderepiek. Een schatting van het tekstverlies met methodes uit de ecodiversiteit”. In: *Spiegel der Letteren* 61.3 (2019), pp. 271–290.

[27] H. Kienhorst. *De handschriften van de Middelnederlandse ridderepiek. Een codicologische beschrijving. Deel 1*. Deventer studieën 9. Sub Rosa, 1988.

[28] E. Kwakkel. *Books Before Print: Electronic Representations of Literary Texts*. Amsterdam University Press, 2018.

[29] E. Marcon. “Practical Estimation of Diversity from Abundance Data”. working paper or preprint. Oct. 2015. URL: https://hal-agroparistech.archives-ouvertes.fr/hal-01212435.

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

[31] A. Orlitsky, A. T. Suresh, and Y. Wu. “Optimal prediction of the number of unseen species”. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.47 (2016), pp. 13283–13288.

[32] A. Pettegree. “The Legion of the Lost. Recovering the Lost Books of Early Modern Europe”. In: *Lost Books. Reconstructing the Print World of Pre-Industrial Europe*. Ed. by F. Bruni and A. Pettegree. Brill, 2016, pp. 1–27.

[33] G. Proot. “Survival Factors of Seventeenth-Century Hand-Press Books Published in the Southern Netherlands: The Importance of Sheet Counts, Sämmelbande and the Role of Institutional Collections”. In: *Lost Books. Reconstructing the Print World of Pre-Industrial Europe*. Ed. by F. Bruni and A. Pettegree. Brill, 2016, pp. 160–201.

[34] G. Proot and L. Egghe. “Estimating Editions on the Basis of Survivals: Printed Programmes of Jesuit Plays in the Provincia Flandro-Belgica before 1773, with a Note on the “Book Historical Law””. In: *The Papers of the Bibliographical Society of America* 102.2 (2008), pp. 149–174.

[35] N. Pyenson and L. Pyenson. “Treating Medieval Manuscripts as Fossils”. In: *Science* 309.5735 (2005), pp. 698–701.

[36] P. L. Shillingsburg. *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge University Press, 2006.

[37] F. Van Oostrom. *Stemmen op schrift. Geschiedenis van de Nederlandse literatuur van het begin tot 1300*. Prometheus, 2006.

[38] J.-P. Wang. “SPECIES: An R Package for Species Richness Estimation”. In: *Journal of Statistical Software* 40.9 (2011), pp. 1–15.

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

[40] R. Wilson. *The lost literature of medieval England*. Methuen & Co, 1952.