Transmission and Survival of Iberian Patristic Texts (3rd–5th Centuries)

Émilie Guidi[1], Théo Moins[1], and Jean-Baptiste Camps[1]

[1] École nationale des chartes, Université PSL, Paris, France

Abstract

This paper analyses the textual transmission of the Church Fathers from the Iberian Peninsula. The corpus is characterised by formal (prose, verse) and generic (sermons, letters, chronicles, epics) heterogeneity. Our computational analyses reveal two contrasting transmission dynamics: prose texts are more numerous but are transmitted by fewer witnesses, sometimes only via their inclusion in medieval collections. Poetic texts, fewer in number, have generated a higher number of witnesses, likely due to their integration in large literary projects.

We model these dynamics using two approaches: **probabilistic unseen species models**, which estimate an upper bound of text and witness survival rates and indicate low corpus diversity and evenness; and **stochastic birth-death models**, which explore cultural evolutionary patterns in text and witness populations. Results suggest a text survival rate below 67% (potentially closer to 20%) and a manuscript survival rate below 10% (possibly under 1%).

Notably, these estimates diverge from prior findings for Medieval French literature, where unseen species and birth-death models yielded similar results. This discrepancy suggests that diachrony – specifically, the broader chronological range of the patristic corpus – plays a key role in shaping transmission outcomes. Our findings also highlight limitations of the birth-death model, particularly in accounting for highly successful texts and in temporal variations in production/destruction rates.

Keywords: Unseen Species, Agent-Based Modelling, Birth-and-Death Process, Cultural Transmission, Philology, Patristics

1 Introduction

1.1 Computational methods for textual transmisssion

In the field of textual history, textual transmission designates the way a text is disseminated and preserved through the copy of material artefacts like manuscripts. This process of transmission is neither linear nor static; it follows an evolutionary dynamic marked by processes of dissemination, selection, variation and extinction. Books are first produced – for instance, in the case of patristic texts we envision here, manuscripts are usually copied in scriptorial workshops linked to libraries or monastic schools, where scribes copy them with a view to preserving them [1]. Their diffusion, whether immediate or spread over the long term, depends on a number of factors, in particular the decision to reproduce certain texts rather than others. Liturgical manuscripts and the major texts of the Fathers of the Church were widely copied. Texts from Classical and Late Antiquity were also favoured, as they were often used in schools to teach reading and rhetoric. Some manuscripts disappeared, whether for human or natural reasons (gradual loss of interest in a work, library fires, wars, etc.), while others were reused, prolonging their existence in new contexts. The manuscript

Émilie Guidi, Théo Moins, and Jean-Baptiste Camps. "Transmission and Survival of Iberian Patristic Texts (3rd–5th Centuries)." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 556–574. https://doi.org/10.63744/WVZDLY7xI2fT.

tradition thus reflects different uses and intentions at different times. It evolves, transforms and adapts according to chance and to the needs and views that societies have of texts.

Traditional manuscript sciences have long been the principal tools used to describe a textual tradition. Palaeography focuses on the evolution of scripts, while codicology focuses on the material aspects of manuscripts. Philology, which notably includes ecdotics and textual criticism, concentrates on the linguistic and literary study of texts, sometimes attempting to reconstruct earlier versions, or even lost versions, of them [19]. In particular, textual criticism aims to examine the variants present in the different witnesses, organise them in a genealogical perspective, and restore the earliest accessible form of a text based on the available evidence. While these disciplines can help to recover certain aspects of a lost manuscript tradition, they are nevertheless unable to grasp the tradition in its entirety.

This is precisely where new approaches come into play, made possible by the development of computational humanities. Recent years have seen the emergence of computational methods applied to the study of textual transmission, with the aim of modelling the dynamics of loss and survival. Two main approaches have emerged. The first, developed as part of the *Forgotten Books* project [16], applies models of unseen species derived from ecology to manuscript traditions, from a probabilistic perspective. The second, carried out by the *ERC-LostMA* project [4], uses stochastic processes of the birth-and-death type, coupled with computer simulations, in the direction pioneered by the work of Weitzman [23] and Cisne [7].

1.2 Contributions

The main objective of our work is to apply these recent approaches to patristic texts produced in the Iberian Peninsula between the mid-3rd century and the mid-5th century. This textual tradition is still unexplored with such methods, which has so far concentrated on ancient texts [23], medieval scientific manuscripts [7], or chivalric narratives [16]. Studying the transmission of a corpus in this pivotal period between Classical Antiquity and the Middle Ages from this perspective remains therefore unprecedented. We can wonder whether this corpus follows similar patterns of transmission and loss than those previously studied, or if the difference in the length of the period envisioned, and the mode of transmission and reception of those texts (particularly confronted with vernacular fictional literature studied in previous projects) will be revealed through computational analysis.

This study aims to model the transmission of the patristic tradition in the Iberian Peninsula by estimating the survival and disappearance rates of texts and manuscript witnesses. It also examines whether distinct dynamics emerge in the processes of conservation, diffusion, and extinction depending on textual form, particularly between prose and poetry. For this purpose, we will employ probabilistic models, especially those of unseen species [16], as well as stochastic processes [4]. The approach relies on the creation of a new database recording all known manuscript sources of the Iberian patristic tradition of texts originating between the 3rd and the 5th century. It also proposes the application, for the first time, of computational methods to this corpus, which involves adapting the existing computational models to the specific features of Late Antiquity, in order to better account for the dynamics specific to this period.

2 Material

2.1 Data collection

The constitution of the corpus involved manual data collection, consisting of an inventory of the texts of the Church Fathers active in the Iberian Peninsula between the 3rd and 5th centuries, as well as the identifiable manuscript witnesses that transmitted them. The selection focused exclusively on the direct tradition, including both preserved manuscripts (complete or fragmentary) and those

whose existence is attested despite their loss (fire, war, etc.), but not the derived works. This corpus covers the geographical area of the Iberian Peninsula, corresponding to the territories of present-day Spain and Portugal. This choice is particularly relevant because the tradition from this period is well documented and has recently been the subject of critical editions [see 8].

It is delimited chronologically by the Councils of Nicaea (325) and Chalcedon (451), in accordance with the boundaries established by the *Clavis Patrum Latinorum* [12]. Authors of Late Antiquity (6^{th}-7^{th} centuries) have been deliberately excluded, since many of their texts have not yet been reliably edited or their manuscript tradition studied in depth. The identification of texts and authors was done mainly on the basis of the *Clavis Patrum Latinorum* [12] [henceforth *CPL*], supplemented by the *Traditio Patrum, Scriptores Hispaniae* [8]. Authors were selected who were either of Iberian origin or who were active in the region. Texts attributed with certainty, but also those described as *dubia* or *spuria*[1], were included, as long as their author's origin was recognised as Iberian. Two anonymous texts have also been included (*CPL*, 789 and 373a), their production context being located on the Iberian peninsula.

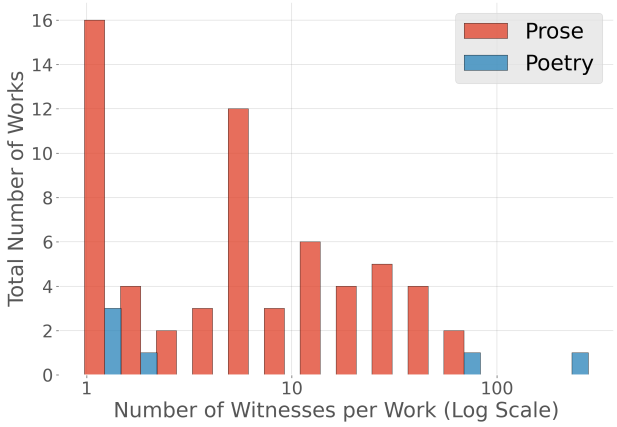
Indirect tradition has been deliberately excluded from our study. In our corpus, it manifests itself in a variety of forms. Some texts are taken up, summarised or adapted in other works. This is particularly true of the *Chronicle* of Hydace, an epitome of which is incorporated into later compositions such as the *Historia Gothorum, Wandalorum et Suevorum* (*CPL* 1204) by Isidore of Seville, or the *Pseudo-Fredegarius*. This phenomenon also applies to poetry, where texts such as those by Prudentius served as models or were taken up in intertextual chains, notably involving Arator's *Apostolic History*, Corippus' *Johannides* or, more recently, Milon de Saint-Amand's *De Sobrietate*. A second case that is also excluded is that of recensions, as shown by the example of Montanus of Toledo, whose *Epistulae* (*CPL*, 1094) circulated within the *Collectio Hispana*. The latter itself gave rise to derivative versions, such as the *Hispana Gallica*.

Excluding indirect tradition and recensions thus allows us to focus on the corpus proper and the diversity of its textual forms, both in prose and versified form. On the prose side, we find a wide variety of genres: letters and treatises, writings related to preaching such as sermons and homilies, texts of confession of faith such as *credo*, foundational texts such as ecclesiastical canons, as well as historiographical texts including historical chronicles. It is worth highlighting a specific feature of the transmission of the texts in our corpus: Late Antique texts circulated in the form of collections [22], structured according to thematic or generic coherence. Some texts owe their preservation to their inclusion in homilies, others to their inclusion in canonical collections. Still others have been preserved as part of vast doctrinally-oriented collections, notably on themes such as Priscillianism and other theological controversies. The Ancients did very little theorising on literary genres, a fact that is particularly evident in prose, which contains many hybrid work, such as the "lettres-traités", that blend correspondence and morale treatise. The text *De similitudine carnis peccati* (*CPL*, 567) by the priest Eutrope is a representative example as it's a letter addressed to Caesaria, a pious woman, combining the personal form of an epistolary exchange with a structured theological argument on the nature of Christ and the incarnation. On the poetic side, the corpus includes verse texts ranging from relatively short poems covering a variety of themes such as panegyric, lyrical, satiric, and didactic pieces, to large-scale works such as epics, which are represented by major figures like Prudentius and Juvencus.

2.2 Corpus analysis

Applying these criteria, we assembled a corpus of $M = 1159$ manuscripts corresponding to $T = 67$ texts by 37 different authors. The oldest manuscripts date from the 6th century, while we set

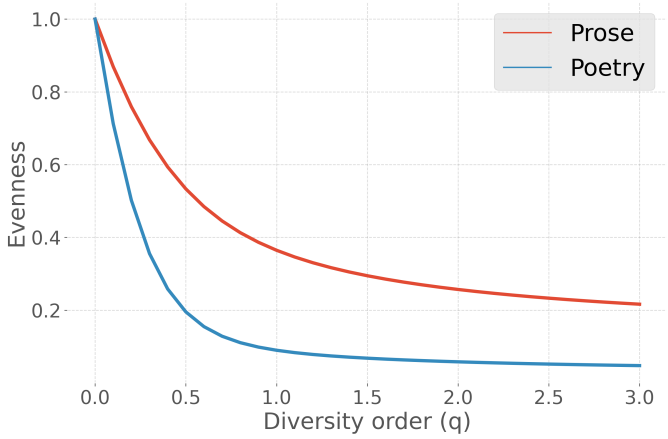
[1] Works considered *dubia* are those whose attribution is questioned, and works considered *spuria* are those whose attribution has been recognised as false.



(a) Abundance data distribution for prose and poetry.

Genre	f_1	f_2	T	M
Poetry	3	1	6	365
Prose	16	4	61	794
Total	19	5	67	1159

(b) Singletons f_1, doubletons f_2, number of texts T and number of manuscript witnesses M for each part of the corpus.



(c) Evenness profiles for prose and poetry texts across orders q

Figure 1: Abundance and evenness comparison for prose and poetic texts.

the 15th century as the upper limit, since in our tradition manuscripts later than the 15th century are often copies based on printed editions, whereas we are primarily interested in witnesses of the manuscript tradition itself. Their provenance spans the whole of Europe.

The dataset is illustrated in Figure 1, and highlights a marked contrast between the two genres. Prose exhibits a significantly greater richness, with 61 distinct texts identified, compared to only 6 for poetry. However, despite this disparity in textual diversity, poetry shows a remarkably high total abundance ($M = 365$ poetry manuscripts), concentrated in a small number of works. In contrast, the 61 prose texts collectively amount to a total abundance of $M = 794$, see Figure 1b. This disproportionate distribution indicates a strong transmission dynamic for a small number of poetic texts, reflecting their cultural or literary prominence in our corpus. To further characterize this distribution, we plot in Figure 1c diversity profiles based on the parameter q, which controls the sensitivity to dominant versus rare elements: at $q = 0$, diversity corresponds to richness (number of distinct texts), while higher values of q progressively emphasize the most abundant texts. Evenness, derived from these profiles, measures how equitably transmission is distributed across texts, with values near 1 indicating balanced distribution and lower values reflecting concentration around a few dominant works. As shown in Figure 1c, the evenness profiles reveal a marked difference between prose and poetry texts. Across all orders q, the curve corresponding to poetry declines more rapidly and reaches significantly lower values than that of prose. This steep drop,

observable even at low values of q, indicates a strong concentration of transmission around a few highly dominant poetic works, to the detriment of the rest of the corpus. In contrast, the prose curve follows a more gradual slope and maintains higher evenness values, suggesting a more balanced distribution among the various texts. How can we explain and model the fact that poetic texts, though fewer in number, were widely transmitted and frequently copied? In particular, how does this specific behaviour influence our analyses of text survival? A first response would be to make two separate studies for poetry and prose, but the small number of poetry samples forces us for now to study only the prose texts for the rest of the study. Results on the entire corpus (prose and poetry) are also given and discussed in Appendix D.

3 Methods

3.1 Unseen Species Model

Recent research demonstrated that the statistical diversity estimators, traditionally commonly used - among other fields - in ecology, can also be successfully applied to cultural heritage [16]. The non-parametric methods, such as unseen species models, make it possible to estimate total richness (i.e., the total number of species, or texts in our case) and diversity of a population, by integrating unobserved elements, and correct the bias induced by the frequent presence of rare species, a major characteristic of our corpus. In particular, the *Forgotten Book* project applied several estimators to a corpus of European medieval chivalric narratives, in several languages, thanks to which the authors were able to estimate a minimal number of unseen (i.e. lost) texts and manuscript witnesses [16]. The methods they introduced to the field of philology were made available in the *Copia* Python package, that has also since been enriched with additional estimators [15].

We choose three estimators used in *Forgotten Books* [16]: Chao1 [5], iChao1 [6] and Jackknife [2]. The Chao1 estimator provides a minimum bound on the original richness based on texts preserved in one or two witnesses (singletons and doubletons), while iChao1 refines this estimate by also including texts present three or four times, thus reducing the risk of underestimation. This approach is particularly well-suited to our corpus, in which 40% of the texts (27 in total) are transmitted by only a very small number of witnesses (one to four witnesses). The Jackknife estimator stands out for its ability to take into account the heterogeneity of the distribution of texts, which is crucial in a corpus where a small number of widely distributed texts, such as Prudentius texts that total almost 300 witnesses, coexist with others that are extremely rare. Unlike Chao1 and iChao1, which rely mainly on the frequency of the rarest texts, Jackknife uses a systematic resampling mechanism that exploits all the frequencies of appearance, providing a nuanced estimate of the original richness[2].

In order to estimate the minimum number of witnesses that would need to be observed in order to register all the unobserved texts in the Iberian patristic tradition, we used the estimator *Minimum Additional Sampling* [16], included in *Copia*, that was used to estimate the survival rates for documents preserving romances and heroic narratives in different linguistic traditions [16]. This estimator allows us to estimate how many additional witnesses would be required for an extended sample to contain at least one occurrence of each original text. It is based on the dynamics of singletons when the sample is enlarged: those already present in the initial sample versus those newly discovered.

On the basis of these richness estimators, we also estimated the survival rate of texts and witnesses: to do this, we converted the estimates of textual and witness richness into survival pro-

[2] Although the ACE estimator is available in the Copia package [15], we do not consider this method here as it requires an arbitrary abundance threshold corresponding to the upper limit for considering a species as rare, and complicates comparison between estimations made on different corpora, that could require the use of different thresholds. Similarly, Egghe's estimator [11], also implemented in Copia but initially designed for printed corpora with a known number of print runs, is not relevant to a manuscript tradition.

portions, by relating the estimated richness to the richness actually observed. We then generated density curves by resampling using the *bootstrap* method, making it possible to visualise the variability of the survival rates for each estimator selected (see below, Figure 8a and Figure 9).

3.2 Modelling the dynamics of text transmission

Alternatively, one can consider a dynamic approach by simulating the temporal evolution of both text and witness populations, in order to reconstruct the evolutionary processes that led to the formation of our current corpus. The real-world dynamics of transmission or extinction of texts were probably influenced by a multiplicity of factors, be they extrinsic (fires, wars, economic crises, etc.) or intrinsic (different fitness of the texts or the witnesses, reception and reader interest, etc.). From the modelling perspective, the core transmission dynamic reduces to a new text appearing with an original witness (the authorial manuscript), and witnesses that can be copied and/or destroyed in time. For each text, this implies two outcomes: survival if there is more copy than destruction, or disappearance.

A primary method for modelling involves employing a birth-death process with constant rates [4; 23]: each witness is modelled as an independent agent, from which copies are made with rate λ per unit time and that can be destroyed with rate μ. A recent update applied to Medieval chivalric literature [4] allows parameters λ and μ to be estimated from observations, in a model with a temporal division into two distinct phases: a period of activity, when texts were actively transmitted and copied in manuscript form, and a period of inactivity when the texts are no longer copied, but witnesses can still be destroyed, and so λ is fixed at 0 but not μ (modelling the period after the appearance of the printing press). Obtaining an estimate of λ and μ allows to deduce an estimate of quantities of interest such as survival rate, extinction probability, topological properties of trees, etc., either by mathematical derivation or by re-simulations of this process with estimated parameters.

Yet, our corpus presents a relatively different type of tradition, particularly regarding chronological parameters, since the texts go back to Late Antiquity. Moreover, while the texts were conceived by authors from the Iberian peninsula, during a limited amount of time (the 3rd to 5th centuries), copies continued to be actively produced throughout Europe during the centuries of the Middle Ages. Therefore, some adaptations are needed with respect to the model used by Camps et al. [4], in order to distinguish a period of both texts and witnesses creation, to a period where no new text appear but witnesses keep being reproduced.

To overcome these obstacles, we suggest a model of transmission divided into three phases:

1. **An innovation phase (201–500)**: This corresponds to Late Antiquity, when patristic authors were actively producing new texts while existing texts were simultaneously being transmitted through manuscript copying. During this period, new independent texts are created while existing manuscript witnesses continue to be copied and potentially destroyed. This phase is characterized by the two parameter λ (copying rate) and μ (destruction rate), but also by a third one, Λ, for new tree creation per unit time. Λ act as an innovation rate, and is not proportional to the average population (contrary to λ and μ), but represents the external appearance of a new text.

2. **A reproduction phase (501–1450)**: This corresponds to the medieval period, when patristic texts were no longer being authored but were actively transmitted throughout Europe through manuscript copying, with scribes reproducing existing texts without creating fundamentally new texts. This phase is marked by the cessation of new text creation ($\Lambda = 0$), while existing witnesses continue to be copied and destroyed according to rates λ and μ respectively.

3. **An inactive phase (1451–1700)**: This phase typically corresponds to the post-medieval

period, particularly after the advent of printing, when manuscript copying was largely abandoned while existing witnesses remained vulnerable to loss and destruction. During this final period, copying activity ceases ($\lambda = 0$), leaving only the destruction rate μ active. Witnesses can still be lost through various factors (wars, fires, material degradation, neglect) but are no longer being reproduced.

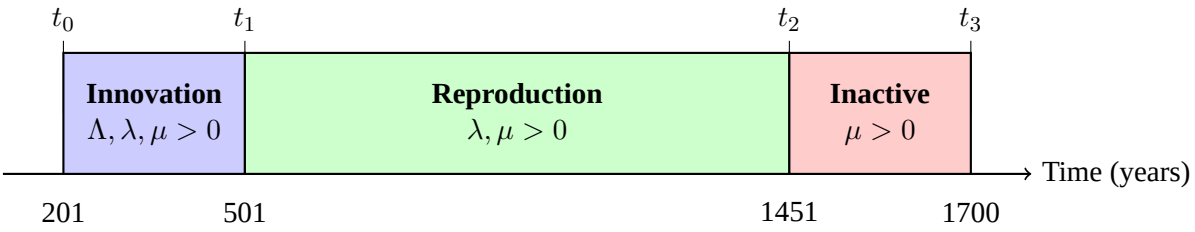


Figure 2: Timeline of the three-phase transmission model with active parameters for each period.

A diagram of the different periods, along with the definition of the time points t_i's, that mark the start/end times for each phase, is summarised in Figure 2. This model allows an explicit calculation of quantities of interest like the richness or the proportion of surviving texts and witnesses at t_3, as a function of the parameters (Λ, λ, μ) and duration of the different phases t_i. The various formulas are given in subsection B.1.

Parameter estimation for (Λ, λ, μ) from observed data constitutes a classical inverse problem: while we can simulate this kind of process from given parameters, we need to infer parameters from observed data. We employ here Simulation-Based Inference (SBI) [9], which simulates the transmission process across various parameter combinations and learns the mapping between parameters and resulting data to identify the most probable parameters generating our observations. Unlike point estimation methods, this Bayesian framework provides full posterior probability densities $p(\Lambda, \lambda, \mu \mid \text{data})$, quantifying uncertainty in parameter estimates and enabling robust statistical inference. Therefore, all the quantities of interest obtained by parameter transformation such as richness will be probabilistic, and so all density of probability will be available for inference, in particular to describe the uncertainty around the estimation (more details in subsection B.2). The method requires defining a prior distribution $p(\Lambda, \lambda, \mu)$ over parameter space: we use constrained uniform priors that enforce biological constraints (e.g., $\lambda > \mu$ for population growth) and computational limits through a maximum population threshold `max_pop` to prevent memory overflow during simulation. The inference pipeline, implemented in the `simmatree` package [18], uses Neural Posterior Estimation [20] to approximate the posterior distribution directly via neural networks.[3] Model validation is performed through posterior predictive checks, where parameters sampled from the posterior are used to simulate new datasets and compare them with original observations to assess model adequacy (see Appendix C).

4 Results

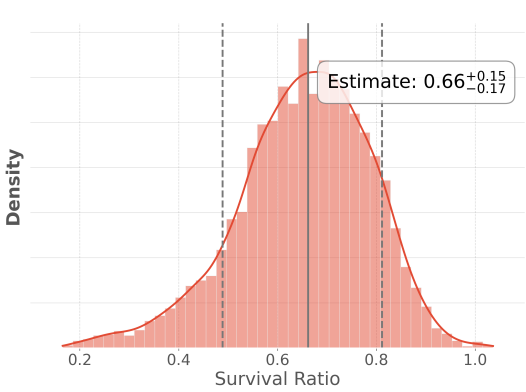
As mentioned above, both methods will be applied to the prose texts of our Iberian patristic corpus, and results on the entire corpus can be found in Appendix D.

Results are given in Figure 3a for text survival rate. The application of unseen species estimators and birth-death modeling to our prose corpus reveals contrasting perspectives on textual loss (Table 1). The unseen species estimators converge on a relatively narrow range for the lower bound of original text richness, with Chao1 suggesting 93 texts, iChao1 indicating 99 texts, and Jackknife estimating 89 texts. These estimates, representing minimum bounds, imply that the 61 observed

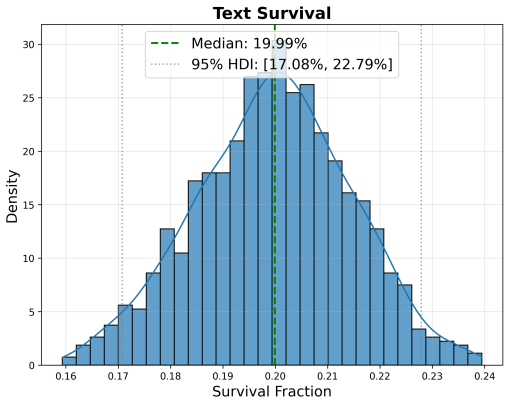
[3] The package also offers multi-round refinement to improve estimation accuracy [10; 20].

Estimator	Text Richness	95% CI
Chao1	93	[64, 173]
iChao1	99	[70, 178]
Jackknife	89	[70, 108]
Birth–Death	313	[220, 338]

Table 1: Estimation of textual richness using Chao1, iChao1, Jackknife, and the one from the Birth–Death model, with 95% confidence intervals (CI) for each estimator (prose texts).



(a) Bootstrap distribution of text survival rates upper bound for iChao1 estimator.



(b) Posterior distribution of text survival rates using the birth–death modelling.

Figure 3: Estimations of text survival rates on the prose corpus, using iChao1 and birth–death forest method.

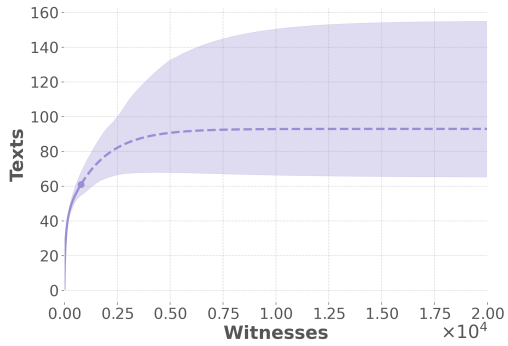
prose texts constitute at most 67% of the original production (see Figure 3a). In contrast, the birth-death model yields a substantially higher estimate of 313 texts for the original corpus. This threefold difference between approaches is noteworthy: while unseen species methods suggest an upper bound of 67% survival based on the iChao1 estimator, the birth-death model's posterior distribution centers around a 20% survival rate (Figure 3a). These divergent estimates reflect the different mathematical frameworks underlying each approach—one providing conservative lower bounds, the other attempting to model the full transmission dynamics.

Looking at Figure 4 and Figure 5a, the disparity between methods becomes even more pronounced when examining manuscript witnesses. The Minimum Additional Sampling approach suggests that current witnesses represent at most 10% of the original manuscript production (Figure 4), but the birth-death model indicates a median survival rate of merely 0.5%, though uncertainty in the model produces a broad distribution extending up to 1.5% (Figure 5a). This uncertainty translates into dramatic variation in absolute numbers: the 95% credible interval spans from 37,000 to over 2 million original witnesses, with a median estimate of approximately 160,000 manuscripts. This suggests a variety of possible scenarios that could lead to the same number of surviving witnesses.

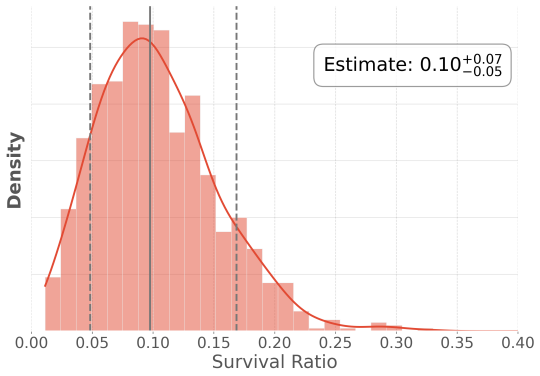
5 Discussion

The previous results deserve to be discussed both from a methodological and a text historical perspective.

From a methodological standpoint, the apparent difference in the estimate provided by unseen species estimators and a birth-death process do not imply, strictly speaking, a contradiction. Since

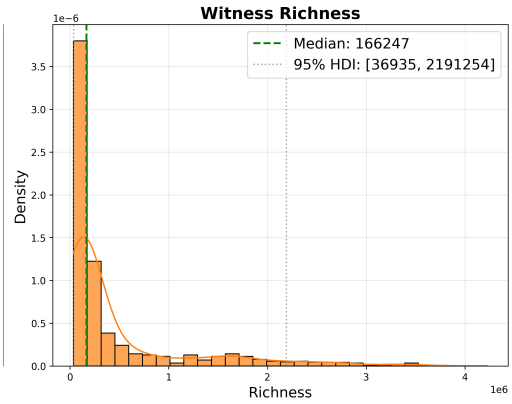


(a) Species accumulation curve. The purple point represents empirical data, the dashed line is the extrapolation curve, and the purple shaded area illustrates the uncertainty range.

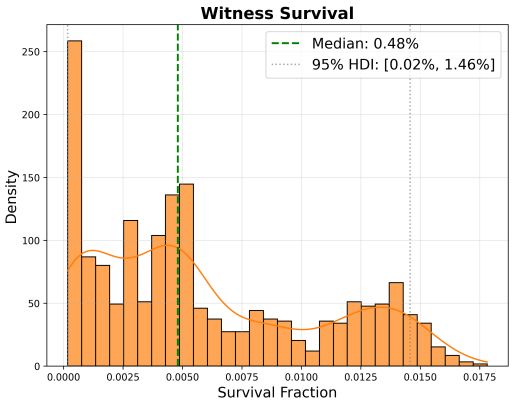


(b) Witnesses survival rates. Estimation and variability using the *bootstrap* procedure. Vertical solid grey line corresponds to the estimate, and dashed lines represent the confidence interval at 95%.

Figure 4: Estimated survival rates by *bootstrap* and witness richness extrapolation for prose corpus



(a) Posterior distribution of witness richness.



(b) Posterior survival rate of witnesses.

Figure 5: Posterior distribution of richness and survival rate for witnesses, using the birth–death forest modelling, for the prose corpus.

said estimators provide a lower bound to richness (and consequently an upper bound to survival), while the birth-death approach provides an estimation of richness (that should consequently respect the bound induced by unseen species), they are actually not in contradiction. It may very well be that the total original richness of texts was superior or equal to 89 texts, while actually being closer to 313. Yet, an obvious limit of the birth-death approach for now is the large uncertainty regarding the survival rate of witnesses (Figure 5a). This uncertainty can have multiple causes: among them, some inadequacies of the birth-death model when compared to empirical data, already observed by Camps et al. [4], that pertain to the distribution of witness per text and to the use of constant rates of copy and destruction. Indeed, the birth-death model induces a distribution of the number of witnesses per texts that decreases exponentially with the number of witnesses, making successful texts with many copies extremely unlikely. Yet, here, we observe that our corpus is very uneven, with a high proportion of one-witness texts, and at the same time a non-zero probability of having many witnesses for a text (see Figure 1). This suggests a power-law behaviour, which is the case in many cultural diffusion phenomena [21]. In particular, the case of poetry, despite the low number of observations, is even worse: half the observations have only one witness, and conversely, the text of Prudence has a number of witnesses 4 times higher than any other text (and 2 orders of magnitude greater than the median value). Integrating other dynamics in modelling could improve the ability of the model to account for empirical data. In particular, modelling the interrelationship between texts, and the process of text creation through derivation from preexisting texts, has recently been shown as a potential way to solve this issue [3]. This would necessitate to include in the dataset derivative texts, such as later recensions or rewritings of our texts.

Moreover, constant rates, though an useful simplification when designing a null model, do not account for extrinsic factors and are not realistic in themselves. It seems really unlikely, from an historical perspective, that the rate of production (or destruction) of books in, say, the 6th century would be the same that in the 15th century. On this aspect, further research should investigate additional refinements to the birth-death approach, that could better account for the distribution of abundance data and for variations in time. Current research is exploring adaptations to the Birth-Death process that integrate variable rates [13], and they could be applied to this corpus in the future.

From a text historic perspective, it is interesting to compare the estimations obtained here with those obtained for Medieval French epics and romances, much later genres (11th-15th c.), of leisure literature in vernacular, marked by a very active transmission process, involving many rewritings [4; 16]. The upper bound of survival rates given by unseen species methods is more optimistic for patristic texts, with up to 70% of survival for texts and 10% for witnesses, while the same method yields values of respectively 55% and 5% for these Medieval French narratives. Even if one were to hypothesise that patristic texts and manuscripts were handled with much more care and reverence than leisure literature in vernacular, these differences still remain hard to explain from an historic perspective, given the substantially longer chronological range involved in the transmission of the Iberian patristic corpus (more than 1,000 years to the appearance of the printing press instead of 400), moreover encompassing such events as the Vandal and Wisigothic migrations, the end of the Western Roman Empire or the Muslim conquest of the Iberian Peninsula, to name but a few. Results obtained using BD processes hinted at 40 to 50% of survival of Old French texts, quite close to the bound provided by unseen species methods, and to 1% of survival for witnesses, while they are closer to respectively 20% and 0.5% for the patristic corpus, this time being significantly more pessimistic than the unseen species results. Since knowledge of the historical context tends to pull into two opposite directions (i.e., greater care for patristic texts and manuscripts, but longer chronological range filled with destructive large-scale events), it seems impossible for now to give a firm advantage to one of these estimate. Yet, this calls for an investigation on the potential effects of diachrony on the results provided both by unseen species

estimators or birth-death processes.

Additionally, the results tend to show significant differences in transmission dynamics for poetic and prose texts. Although our corpus contains relatively few poetic texts, they occupy an important place in the manuscript tradition, as shown by the significant number of witnesses associated with them. This success can be explained on several levels. The poetic form, applied to works of a theological nature, introduces a stylistic and aesthetic refinement that contrasts with the sobriety of prose. Poetry plays an ornamental role, while at the same time having an edifying aim: the aesthetic pleasure it provides makes it easier to convey doctrinal precepts in a more attractive way. Although poetic texts generally enjoyed wide circulation, their diversity was limited compared to that of prose. This low diversity can be explained by the presence of a few texts with a very high number of witnesses, indicative of their success. Juvencus and Prudentius, in particular, were able to combine Christian tradition with ancient literary heritage, notably through poetic projects inspired by classical epic while conveying Christian doctrinal content. Juvencus is the author of the first Christian epic, the *Evangeliorum libri IV*(*CPL* 1385), while Prudentius developed an allegorical epic (*CPL* 1437-1441). This originality might have favoured their reception and dissemination in scholarly circles, especially in medieval schools where they were widely studied. Their prominence within the corpus is striking: together, these two works account for more than 300 of the 365 extant witnesses of Christian Latin poetry, creating a marked imbalance in the distribution. Moreover, in terms of the entire patristic tradition, they represent the two most widely transmitted texts. This observation also explains the different evenness profiles observed in Figure 1c.

As for prose texts, their predominance in our corpus is primarily explained by the very nature of these works. Unlike poetry, which is often reserved for more elaborate literary projects, prose lends itself more to short, functional forms such as sermons, homilies and correspondence. These genres were widely used in ecclesiastical circles in Late Antiquity, and their abundance was due less to literary prestige than to regular use in teaching, preaching or liturgy. Their short form also facilitated their inclusion in medieval collections, such as the homilies, bearing witness to a collective transmission. It is maybe this grouping effort that explains why short and sometimes less literary texts were able to survive.

Code and materials availability

The code and datasets used in this paper are available on Zenodo, `https://doi.org/10.5281/zenodo.17456880`.

Acknowledgements


Funded by the European Union (ERC, LostMA, 101117408). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

This work has received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by ANR with the references ANR-10-IDEX-0001.

References

[1] Bischoff, Bernard. *Paléographie de l'Antiquité romaine et du Moyen Age occidental*. Grands manuels Picard. Paris: Picard, 1985.

[2] Burnham, K. P. and Overton, W. S. “Estimation of the size of a closed population when capture probabilities vary among animals”. In: *Biometrika 65*, no. 3 (1978), pp. 625–633. DOI: 10.1093/biomet/65.3.625.

[3] Camps, Jean-Baptiste, Christensen, Kelly, Godreau, Ulysse, and Moins, Théo. “One tree to Yule them all? Reflexions on intertextuality and text transmission”. In: *Digital Humanities 2025 (DH2025)*. 2025.

[4] Camps, Jean-Baptiste, Randon-Furling, Julien, and Godreau, Ulysse. “On the transmission of texts: written cultures as complex systems”. 2025. arXiv: 2505.19246 [physics.soc-ph]. URL: https://arxiv.org/abs/2505.19246.

[5] Chao, Anne. “Non-parametric estimation of the classes in a population”. In: *Scandinavian Journal of Statistics 11*, no. 4 (1984), pp. 265–270. DOI: 10.2307/4615964.

[6] Chiu, Chun-Huo, Wang, Yi-Ting, Walther, Bruno A, and Chao, Anne. “An Improved Nonparametric Lower Bound of Species Richness via a Modified Good–Turing Frequency Formula”. In: *Biometrics 70*, no. 3 (2014), pp. 671–682. DOI: 10.1111/biom.12200.

[7] Cisne, John L. “How science survived: medieval manuscripts “demography” and classic texts extinction”. In: *Science 307*, no. 5713 (2005), pp. 1305–1307. DOI: 10.1126/science.1104718.

[8] Emanuela Colombi, edited by. *Traditio Patrum I. Scriptores Hispanae*. Vol. 4. *Corpus Christianorum. Claves - Subsidia*. Turnhout: Brepols, 2015.

[9] Cranmer, Kyle, Brehmer, Johann, and Louppe, Gilles. “The frontier of simulation-based inference”. In: *Proceedings of the National Academy of Sciences 117*, no. 48 (2020), pp. 30055–30062.

[10] Deistler, Michael, Goncalves, Pedro J., and Macke, Jakob H. “Truncated proposals for scalable and hassle-free simulation-based inference”. In: (2022). arXiv: 2210.04815 [stat.ML]. URL: https://arxiv.org/abs/2210.04815.

[11] Egghe, Léo and Proot, G. “The estimation of the number of lost multi-copy documents: A new type of informetrics theory”. In: *Journal of Informetrics 1*, no. 3 (2007), pp. 257–268. DOI: 10.1016/j.joi.2007.02.001.

[12] Gaar, Emilien. *Clavis patrum latinorum: qua in Corpus christianorum edendum optimas quasque scriptorum recensiones a Tertulliano ad Bedam*, ed. by Dekkers Eloi. 3e édition augmentée. *Corpus christianorum. Series Latina*. *Steenbrugis : in Abbatia Sancti Petri* ; Turnhout: Brepols, 1995.

[13] Godreau, Ulysse, Moins, Théo, Christensen, Kelly, and Camps, Jean-Baptiste. “Why do older books survive (sometimes)? Modelling the time distribution of manuscripts with a birth-death approach”. In: *Computational Humanities Research 2025 (CHR 2025)*. 2025.

[14] Kendall, David G. “On the generalized” birth-and-death” process”. In: *The annals of mathematical statistics* (1948), pp. 1–15.

[15] Kestemont, Mike and Kardsorp, Folgert. “Copia (Python package)”. 2025. URL: https://github.com/mikekestemont/copia.

[16] Kestemont, Mike, Karsdorp, Folgert, Bruijn, E. de, Driscoll, M., Kapitan, Katarzyna A., Ó. Macháin, P., Sawyer, D., Sleiderink, R., and Chao, Anne. “Forgotten books: The application of unseen species models to the survival of culture”. In: *Science 375* (2022), pp. 765–769. DOI: 10.1126/science.abl7655.

[17] Moins, Théo, Arbel, Julyan, Girard, Stéphane, and Dutfoy, Anne. “Reparameterization of extreme value framework for improved Bayesian workflow”. In: *Computational Statistics & Data Analysis 187* (2023), p. 107807.

[18] Moins, Théo, Christensen, Kelly, Camps, Jean-Baptiste, and Godreau, Ulysse. “LostMa-ERC/simMAtree: CHR 2025 Release”. Version v0.2.1. Oct. 2025. DOI: 10.5281/zenodo.17425020. URL: https://doi.org/10.5281/zenodo.17425020.

[19] Muzerelle, Denis. *Vocabulaire codicologique: répertoire méthodique des termes français relatifs aux manuscrits*, ed. by Cemi. Paris: IRHT, 1985. URL: https://codicologia.irht.cnrs.fr/.

[20] Papamakarios, George and Murray, Iain. “Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation”. In: *29* (2016), ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett.

[21] Pinto, M.A. Carla, Lopes, António M., and Machado Tenreiro, J.A. “A review of power laws in real life phenomena”. In: *Communications in Nonlinear Science and Numerical Simulation 17*, no. 9 (2012), pp. 3558–3578.

[22] Stéphane, Gioanni and Benoît, Grévin. *L’Antiquité tardive dans les collections médiévales : textes et représentations VIe-XIVe siècle*. Collections de l’Ecole Française de Rome. Rome: Ecole Française de Rome, 2008.

[23] Weitzman, Michael P. “The Evolution of Manuscript Traditions”. In: *Journal of the Royal Statistical Society. Series A (General) 150*, no. 4 (1987), pp. 287–308.

A Structure and Content of the Database

The database lists all the known witnesses and texts of the Fathers of the Church active in the Iberian Peninsula between the 3rd and 5th centuries. Each entry corresponds to a witness and includes the following information: siglum, shelfmark, date, geographical origin, and foliation. The textual content transmitted in the witness is also specified, along with the corresponding reference in the *Corpus Patrum Latinorum*, a summary of the text, its place of composition when known, its form (prose or verse), its genre (homily, historical chronicle, letter, essay, etc.), the author’s name, and the century in which he was active. However, three particular features of the database’s structure deserve to be highlighted. First, when a text is anonymous, it is referenced using the format `an_nomtexte`, where the prefix `an_` explicitly indicates the absence of a precise attribution to a known author. Second, for texts with doubtful or incorrect attribution, the database uses a prefix `-s` followed by a number, and then the name of the presumed author. For example, a text with uncertain attribution to Severus would be coded as `s1_Severus`, where `s1` indicates that it is the first text presumed to be by this author. Finally, some witnesses appear under artificial shelfmarks: philologists who have produced the critical edition of a text mention the existence of multiple witnesses that ensured its transmission, but without providing a precise inventory. To ensure completeness and consistency, these witnesses have been retained in the database under standardised shelfmarks, formatted as `manuscript_` followed by the first letters of the concerned text, and then a number assigned according to the count of unidentified witnesses. For example, regarding Priscillian and his *Canones in Pauli apostoli epistulas a Peregrino episcopo emendati*(*CPL*, 786), there are 21 attested witnesses, but only the shelfmarks of 12 are known. Consequently, the remaining nine are recorded using standardized codes such as `manuscript_canones_13`, `manuscript_canones_14`, and so forth.

B Mathematical details on the stochastic process

B.1 Derivation of quantities of interest

Our model can be seen as birth–death process with parameters $(\lambda(t), \mu(t))$ and an innovation rate $\Lambda(t)$, such that

$$(\Lambda(t), \lambda(t), \mu(t)) = \begin{cases} (\Lambda, \lambda, \mu) & \text{if } t_0 \leq t \leq t_1 \\ (0, \lambda, \mu) & \text{if } t_1 \leq t \leq t_2 \\ (0, 0, \mu) & \text{if } t_2 \leq t \leq t_3 \end{cases}$$

The mean number of manuscript witnesses $M(t)$ at a given time t is governed by the following differential equation:

$$\frac{\mathrm{d}M}{\mathrm{d}t}(t) = (\lambda(t) - \mu(t))\, M(t) + \Lambda(t),$$

which leads to the following analytical expression:

$$M(t) = \begin{cases} \left(M(t_0) + \frac{\Lambda}{\lambda-\mu}\right) e^{(\lambda-\mu)t} - \frac{\Lambda}{\lambda-\mu} & \text{if } t_0 \leq t \leq t_1, \\ M(t_1) e^{(\lambda-\mu)(t-t_1)} & \text{if } t_1 \leq t \leq t_2, \\ M(t_2) e^{-\mu(t-t_2)} & \text{if } t_2 \leq t \leq t_3. \end{cases}$$

The cumulative sum $M_{\text{cum}}(t)$ of all the witnesses produced at time t verifies the equation

$$\frac{\mathrm{d}M_{\text{cum}}}{\mathrm{d}t}(t) = \lambda(t) M(t) + \Lambda(t),$$

which leads to the formula

$$M_{\text{cum}}(t) = \begin{cases} \Lambda t + \dfrac{\lambda}{\lambda-\mu}\left(M(t) - M(t_0) - \Lambda t\right) & \text{if } t_0 \leq t \leq t_1, \\ \Lambda t_1 + \dfrac{\lambda}{\lambda-\mu}\left(M(t) - M(t_0) - \Lambda t_1\right) & \text{if } t_1 \leq t \leq t_2, \\ \Lambda t_1 + \dfrac{\lambda}{\lambda-\mu}\left(M(t_2) - M(t_0) - \Lambda t_1\right) & \text{if } t_2 \leq t \leq t_3. \end{cases}$$

Note that $M_{\text{cum}}(t)$ is constant for $t_2 \leq t \leq t_3$, as no new witnesses are produced in this period.

For the mean number of texts $T(t)$ at time t, we introduce the probability $p_0(t_a, t_e)$ of a text to be extinct at time t_e with t_a the ending time of the active phase. In other words, $(\lambda, \mu) > 0$ for $0 \leq t \leq t_a$ and $\lambda = 0$ for $t_a \leq t \leq t_e$. This probability can be written as [4; 14]:

$$\begin{cases} p_0(t_a) & := p_0(t_a, t_a) = \dfrac{\mu\left(e^{(\lambda-\mu)t_a} - 1\right)}{\lambda e^{(\lambda-\mu)t_a} - \mu} \\ p_0(t_a, t_e) & := p_0(t_a) + \left(1 - e^{-\mu(t_e - t_a)}\right) \dfrac{(1 - p_0(t_a))(1 - \frac{\lambda}{\mu} p_0(t_a))}{1 - \frac{\lambda}{\mu} p_0(t_a)\left(1 - e^{-\mu(t_e - t_a)}\right)} \end{cases}$$

Using this expression, we obtain

$$T(t) = \begin{cases} \displaystyle\int_0^t \Lambda\left(1 - p_0(t-\tau)\right) \mathrm{d}\tau = \frac{\Lambda}{\lambda} \log\left(\frac{\lambda e^{(\lambda-\mu)t} - \mu}{\lambda - \mu}\right) & \text{if } t_0 \leq t \leq t_1, \\ \displaystyle\int_0^{t_1} \Lambda\left(1 - p_0(t-\tau)\right) \mathrm{d}\tau = \frac{\Lambda}{\lambda} \log\left(\frac{\lambda e^{(\lambda-\mu)t} - \mu}{\lambda e^{(\lambda-\mu)(t-t_1)} - \mu}\right) & \text{if } t_1 \leq t \leq t_2, \\ \displaystyle\int_0^{t_1} \Lambda\left(1 - p_0(t_2 - \tau, t-\tau)\right) \mathrm{d}\tau = \frac{\Lambda}{\lambda} \log\left(\frac{\lambda e^{\lambda t_2 - \mu t} + \lambda\left(1 - e^{-\mu(t-t_2)}\right) - \mu}{\lambda e^{\lambda(t_2-t_1) - \mu(t-t_1)} + \lambda\left(1 - e^{-\mu(t-t_2)}\right) - \mu}\right) & \text{if } t_2 \leq t \leq t_3. \end{cases}$$

The cumulative sum of tree produced in time is simply $T_{\text{cum}}(t) = \Lambda t_1$. From this result, we obtain our two quantities of interest, the theoretical fractions of surviving texts and witnesses at time t_3, as a function of (Λ, λ, μ):

$$\frac{M(t_3)}{M_{\text{cum}}(t_3)}, \quad \text{and} \quad \frac{T(t_3)}{T_{\text{cum}}(t_3)}.$$

B.2 Definition of the constrained prior

The prior distribution corresponds to the distribution of (Λ, λ, μ) that model our knowledge before observing abundance data. A first approach would be to consider that we don't have any external information on parameters, and so we consider uniform distributions for the prior (despite uniformity is not synonym of non-informativeness [17]), with an upper bound sufficiently large. However, this will suffer of severe inefficiency, as most simulations produced by parameters drawn uniformly on the cube will produce either an empty population (typically when $\lambda < \mu$) or an exploding one (typically when $\lambda - \mu$ is too large) So instead, we add constraints to restrict the uniform distribution on a zone where

$$\begin{cases} \lambda > \mu, \\ M(t_2) < \texttt{max\_pop}, \\ M(t_3) > 1, \end{cases}$$

using the calculation of $M(t)$ as a function of (Λ, λ, μ) of the previous section.

C Details on (Λ, λ, μ) estimation for the Iberian Patristic Prose Corpus

Figure 6 shows that our approach successfully converged to a well-defined posterior distribution for the three-phase birth-death model parameters (Λ, λ, μ) when applied to the prose corpus. The figure presents a pairplot comparison between the prior and posterior distributions, illustrating how the data has informed parameter estimation. In particular, the median values for the three parameters here are $\hat{\Lambda} \approx 2.6.10^{-1}$, $\hat{\lambda} \approx 4.7.10^{-3}$, and $\hat{\mu} \approx 3.7.10^{-3}$).

Figure 7 presents posterior predictive checks based on 500 simulations drawn from the posterior distribution. These checks evaluate the model's ability to reproduce key summary statistics of the observed data by comparing the empirical values (shown as vertical lines) with the distribution of values obtained from posterior predictive simulations. The majority of summary statistics show satisfactory agreement between the model predictions and observed data. The model successfully captures the number of texts, witnesses, the maximum and median number of observations per text. This indicates that the three-phase birth-death process provides a reasonable approximation of the core transmission dynamics for prose texts in the Iberian patristic tradition. A small discrepancy emerges in the number of texts with one witness, which suggests that while the birth-death model captures the general transmission patterns, certain aspects of the empirical distribution remain inadequately modelled.

D Results on the entire corpus

First, applying unseen species estimators to the corpus of patristic texts from the Iberian Peninsula enables us to assess both the minimal number of texts originally produced and an upper bound of the proportion currently preserved. These results (Figure 8a) suggest that the initial Iberian patristic tradition comprised at least one hundred distinct texts. The bootstrap method enables an evaluation of the variability in the estimated survival rates. For instance, the density curves resulting from the iChao1 procedure indicate that 66% of the texts have likely survived (Figure 8a). This trend

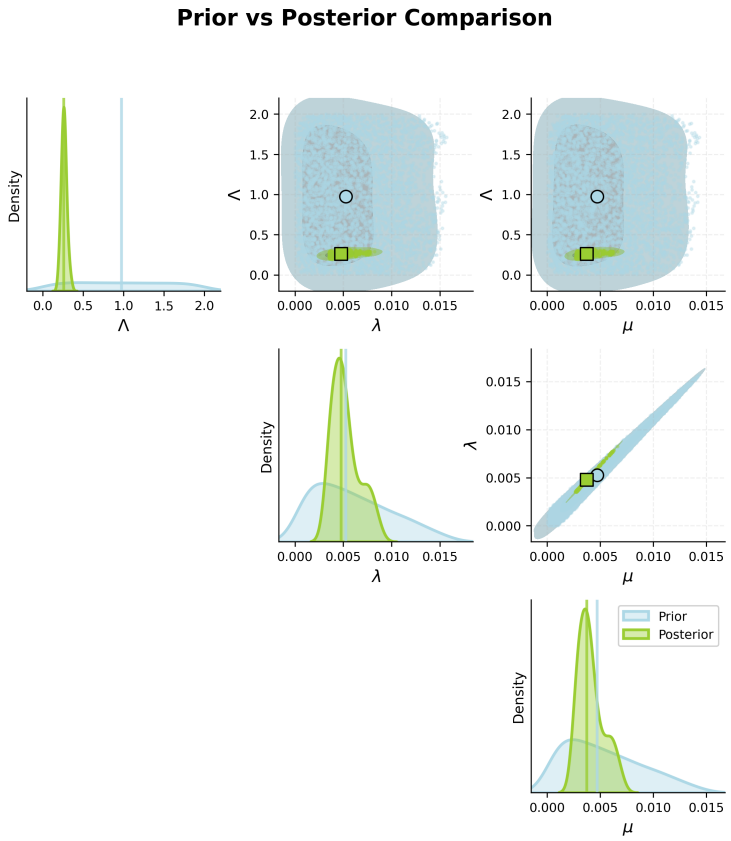


Figure 6: Pairplot comparing prior and posterior distribution of the parameters (Λ, λ, μ) for the prose dataset.

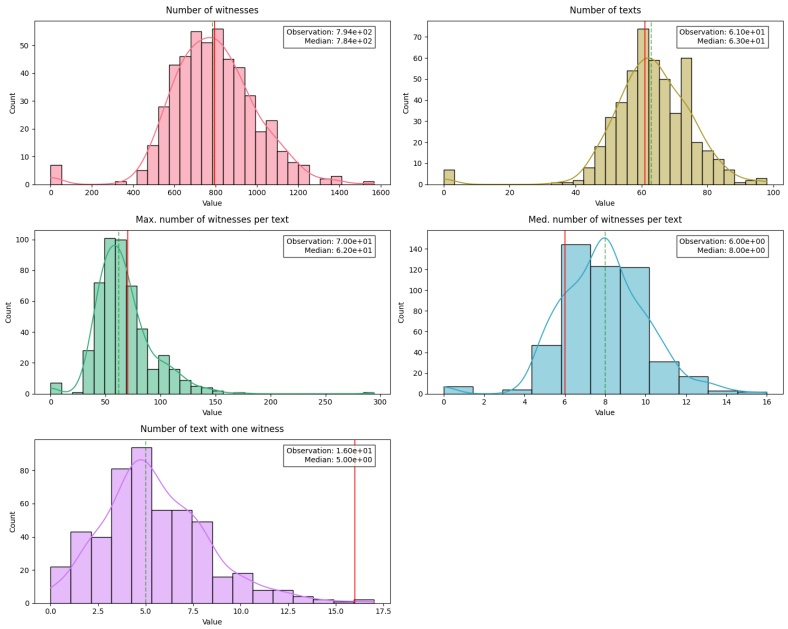
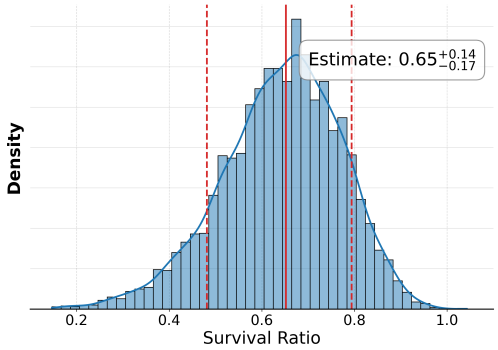
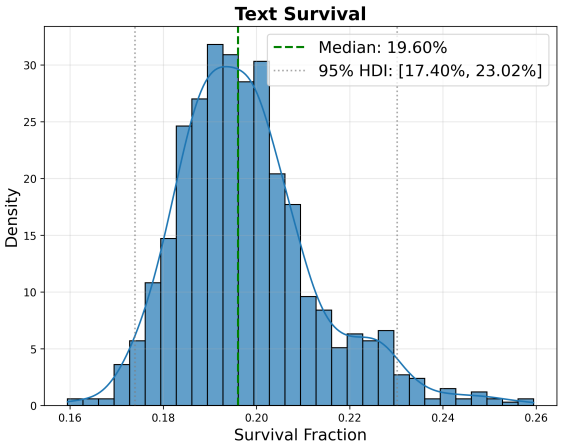


Figure 7: Posterior predictive checks on 500 summary statistics, drawn using parameters samples from the posterior distribution, using the prose corpus.

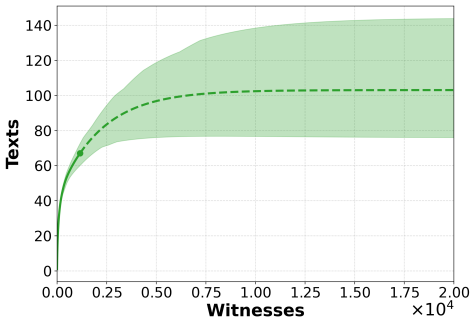


(a) Bootstrap distribution of text survival rates upper bound for iChao1 estimator.

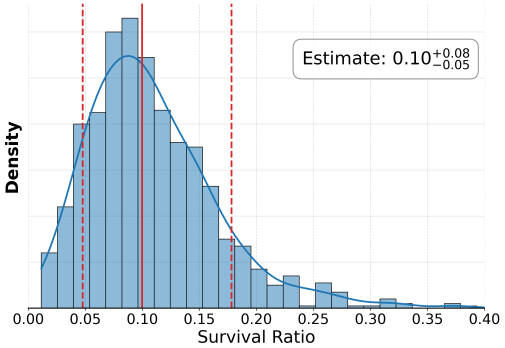


(b) Posterior distribution of text survival rates using the birth–death modelling.

Figure 8: Estimations of text survival rates on the entire corpus (poetry and prose), using iChao1 and birth-death method.

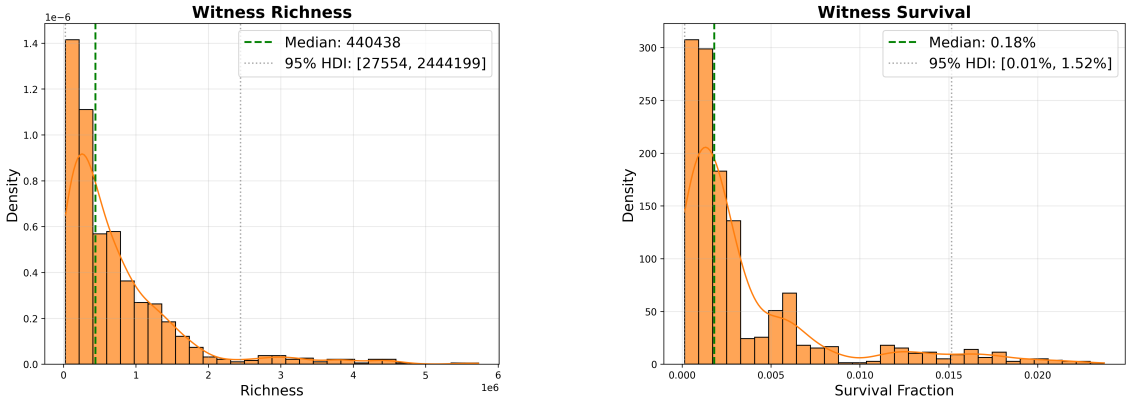


(a) Species accumulation curve. The green point represents empirical data, the dashed line is the extrapolation curve, and the green shaded area illustrates the uncertainty range.



(b) Manuscript witnesses survival rates. Estimation and variability using the *bootstrap* procedure. Line in red correspond to the estimate, and dashed line represent the confidence interval at 95%.

Figure 9: Estimated survival rates by *bootstrap* and witnesses richness extrapolation on the entire corpus (poetry and prose)



(a) Posterior distribution of witness richness.

(b) Posterior survival rate of witnesses.

Figure 10: Posterior distribution of richness and survival rate for witnesses, using the birth–death forest modelling, for the entire corpus (poetry and prose).

suggests that two thirds of the patristic texts produced in the Iberian Peninsula could represent a maximal bound on what has been preserved, thereby moderating the extent of textual loss. To complete this finding, we compare with richness estimation from the stochastic process modelling. Here, contrary to the unseen species results, this estimation is not a lower bound but an estimation of richness supposing a birth–death model. Still, the estimation around 300 stays in the same order of magnitude of all unseen species method, while it remains higher than the highest probable lower of each estimator. Therefore, the results of our modelling do not contradict those of unseen species methods.

For the manuscript witnesses, the application of the Minimum Sampling estimator allowed us to establish a lower bound for the original richness, estimated at approximately 14774 witnesses. The survival rate analysis, based on a bootstrap procedure, indicates a preservation of approximately 10% of the witnesses, with a confidence interval ranging from 4% to 18% (Figure 9b). This result highlights a significant loss in the transmission of witnesses which is greater than that observed for the texts themselves. However, this method may produce by design a lower bound that can be very loose, as it consists in completing the dataset only with the minimum number of samples, itself based on a lower bound of the number of texts. This is illustrated by a significant drop in the survival rate for witnesses for the birth–death method, see Figure 10a: the posterior credible interval at 95% for the survival rate of witness is $[0.01\%, 1.52\%]$, which is far from the 10% rate produced by the minimum sample method. However, using the estimations of the stochastic process point of view reveals difficulties in modelling the dynamics of the entire corpus using our stochastic process approach. Indeed, despite a posterior distribution that seems to be refocused on fairly probable parameter values, the posterior predictive checks (Figure 11) show that none of the replications that can be carried out on the basis of these probable parameters provide abundance data with enough one-text witnesses, and a maximum number of one-text witnesses that is also insufficient to be up to Prudentius.

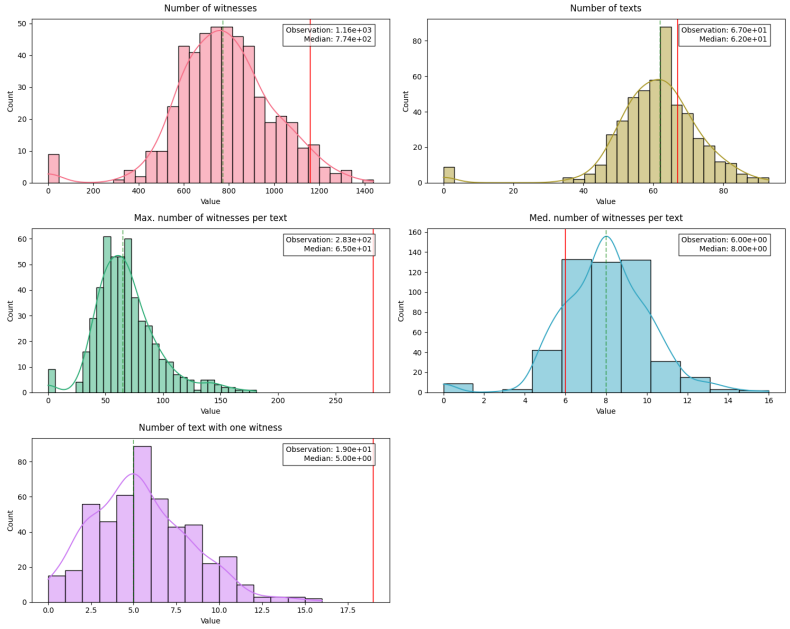


Figure 11: Posterior predictive checks on 500 summary statistics, drawn using parameters samples from the posterior distribution, using the entire corpus.