Recommender systems for fossil community distribution modelling

Zliobaite, Indre

2022-08

Zliobaite , I 2022 , ' Recommender systems for fossil community distribution modelling ' , Methods in Ecology and Evolution , vol. 13 , no. 8 , pp. 1690-1706 . https://doi.org/10.1111/2041-210X.13916

http://hdl.handle.net/10138/352561
https://doi.org/10.1111/2041-210X.13916

Received: 7 June 2021 | Accepted: 10 May 2022

DOI: 10.1111/2041-210X.13916

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Recommender systems for fossil community distribution modelling

Indrė Žliobaitė

University of Helsinki, Helsinki, Finland

Correspondence
Indrė Žliobaitė
Email: indre.zliobaite@helsinki.fi

Funding information
Suomen Akatemia, Grant/Award Number: 314803 and 341623

Handling Editor: Tiago Quental

Abstract

1. We propose to leverage recommender systems from machine learning to build large-scale community distribution models for the mammalian fossil record. Recommender systems are behind most online life today, from shopping to news personalisation, online dating, or the selection of study programmes or fastest routes. Many recommender systems work by predicting user preferences from items that occur together in user profiles. Technically, this setting closely resembles co-occurrence of species in natural environments.
2. Here we frame community distribution modelling as a recommender systems task, tailor existing recommender techniques for this purpose and propose optimisation criteria for fitting the models in the ecological context. The predictive power comes from species co-occurrences.
3. We demonstrate the potential of this approach for analysing past ecosystems on a case study of Miocene fossil sites in Europe, where we use the proposed community distribution modelling for reconstructing companionships and relative abundances of large mammals.
4. The proposed approach to community distribution modelling, although not climatically explicit, can help to reconstruct past ecosystems and analyse their structure and dynamics over time and space. It also allows, even coarsely, to predict relative abundances of fossil species from presence–absence data. More generally, the proposed perspective is a means for analysis of fossil communities and the relationships between their ecological contexts.

KEYWORDS
mammal communities, NOW database, relative abundances, species distribution modeling

1 | INTRODUCTION

Concerns about incompleteness of the fossil record have long been at the forefront of palaeontology research. In the Origins, Charles Darwin dedicated more than a full chapter to imperfection of the geological record (Darwin, 1859). The fossil record known in Darwin's day represented a very small part of potentially available material. With over a century of active collection and documentation of fossil evidence, the global fossil record is now broad and abundant and includes increasingly rich contextual information available via fossil databases (Uhen et al., 2013). If many fragments that represent different circumstances are available, we can hope to

reconstruct ecosystems of the past drawing on overlaps between those fragments.

Extrapolating from overlaps is how modern recommender systems work (Ricci et al., 2011). Recommender systems are behind most of online life today, ranging from dating applications, book recommendations on Amazon, or movies on Netflix to personalised learning, personalised banking, or personalised health. A typical recommender system scenario is that a large list of users is available, where each user has rated, watched or purchased some small number of items. The goal is to predict which other items the user would prefer.

The recommender systems approach naturally lends itself to the analysis of fossil communities if we consider fossil sites as users and fossil taxa as items. One can think of species as components of an ecosystem in a similar way as movies can be components of a user profile. Typical data behind recommender systems are very sparse, as are fossil occurrence data. A single user can realistically watch only a small fraction of the movies that have ever been produced. Similarly, each fossil assemblage represents only a small fraction of all species that have ever lived. As the number of transactions can vary from user to user, similarly the species diversity can vary from site to site. The goal is to predict the preferences of taxa for palaeoenvironments.

One of the key features of the data behind recommender systems is the uncertainty of absence. In the user modelling setting, the presence of a transaction typically signals that the user liked the product, but the absence of a transaction might mean either that the user did not like the product or that he or she has never come across it. Similarly, in the fossil world, the absence of a species at a site may mean that a species was absent from that environment or that it lived there but has not (yet) been found as a fossil.

The uncertainty of absence in ecology is of two main kinds corresponding to presence-background and presence–absence data (Wang & Stone, 2019). The former comes from opportunistic surveys where the presence of species has been recorded but there are no records of absence. The latter comes from systematic surveys in which the confidence about absences is comparable to the confidence about presences. Fossil data are more like the latter. The bulk of the data in curated fossil databases record intensively sampled sites where there is some confidence about the absence. At the same time, absences remain uncertain, since fossilisation is a very unlikely event and even if the site has been well sampled, no representatives of a species that was part of a faunal community in the past may have fossilised.

We propose a methodology for analysing distribution of fossil species and their ecological contexts by leveraging recommender system techniques from machine learning. The computational task is to generalise over a very large and extremely sparse fossil record by modelling preferences of taxa for palaeoenvironments. This will not recreate species that have never been found as fossils but can help to identify viable fossil communities and estimate their completeness and relative abundances. This approach is intended to work with heterogeneous data coming from multiple sources and prior rarefaction or subsampling is not required. Apart from providing proxies for analysing the dynamics of past ecosystems, this approach can offer insights into basic research questions, such as what characterises a viable ecosystem over long times. With a case study of Middle Miocene fossil sites in Europe, we illustrate how such modelling can be used to reconstruct ecological contexts and relative abundances of large mammals.

2 | THE PROPOSED RECOMMENDER SYSTEMS APPROACH

Our task is to quantify the preferences of taxa towards different environments. The classical approach to this task in ecology is species distribution modelling.

2.1 | Rationale

Species distribution modelling, also termed as environmental or ecological niche modelling, aims to predict occurrences of species based on climate or other environmental characteristics (Elith & Leathwick, 2009). Joint species distribution methodology (Pollock et al., 2014; Tikhonov et al., 2020) models species occurrences in relation to environmental characteristics as well as to each other. The mainstream species distribution modelling in ecology focuses on one species at a time, even if co-occurrences are taken as explanatory variables. Multi-species distribution models (Dunstan et al., 2011; Hui et al., 2015) predict responses of multiple species to environmental gradients. A parallel line of research, community composition modelling assesses imperfection of observations as a function of environmental characteristics (Beasley & Maher, 2019). All these approaches primarily rely on climatic data as explanatory variables.

The traditional species distribution modelling has been applied to fossil taxa (Myers et al., 2015; Varela et al., 2011). In those settings, climatic data can come either from climate model simulations or be predicted from morphological, chemical or taxonomic characteristics of the sediments or fossils themselves. One of the main challenges is that localised climatic data of high resolution are not realistically available further than for a very recent past.

To the best of our knowledge, no joint species distribution models, which would consider the distribution of multiple species along with their environmental contexts, have yet been reported for the fossil record. Recently, we attempted an ecometric species distribution modelling of early humans using functional traits of fossil mammals (Saarinen et al., 2021), which is yet another direction close to the traditional species distribution modelling. While not explicitly modelling species distribution or relative abundances. Toth et al. (2019) analysed co-occurrence patterns using a climatic context over the Late Quaternary.

Our proposal is to leverage information of species companionships for modelling probabilities of occurrence. This approach relies on the assumption that species co-occurrence is not random in fossil

communities, that is, those taxa that sometimes occur together are likely to occur together again. While the approach inherently takes into account species interactions, such as competitive exclusion, if any (Blanchet et al., 2020), the main premise for co-occurrence is considered to be the affinity towards similar environmental conditions, climatically or otherwise. The approach does not use climatic estimates explicitly.

The proposed approach can be seen as species distribution modelling via community affiliation, or community distribution modelling, in brief. This is not meant to replace the traditional species distribution modelling, rather to complement it where possible. The approach primarily aims at analyses of large spatial or temporal scales where high-resolution environmental data are not easily or at all available, analysing fossil communities is the primary intended application.

What can we do with such a model in the fossil world? Collaborative filtering, the technique that we employ, produces preference estimates for any taxa to occur at any site. Since estimates come from a single model for many taxa and sites, they are directly comparable across those taxa and sites. Model outputs then be used for three potential purposes: analysing community composition, predicting relative abundances within communities and assisting in data curation.

The model can be used to construct community trees, which would highlight which taxa are more closely related to each other in terms of contexts of their occurrence (companionships). This way we can distinguish occurring together from occurring in similar companionships (while not necessarily often together). We show examples of such analysis with a case study of fossil mammals.

From the curatorial perspective, community distribution models could also be used to assess the quality of the record, flag potentially missing taxa at sites, as well as highlight potential issues with taxonomic identifications. Given preference scores from the model, one could ask what are the species that most likely have occurred at site X in addition to those already found there as fossils? Are there any species that would be highly likely to have been misidentified at site X (present at a site but a model gives a low probability for that species)? While never intended to be used for automated corrections, such predictions could draw attention of data curators to cases for manual inspection. Estimating the number of species that have not been observed is an orthogonal challenge (May, 1988) for which many statistical solutions already exist (Alroy, 2000; Chao et al., 2015; Connolly & Miller, 2001; Foote, 2000, 2016; Raup, 1975). Since common species define major patterns of ecosystems (Jernvall & Fortelius, 2002), for functional analysis of communities the most representative taxa will suffice as long as we can reason about their relative abundances (Vermeij & Herbert, 2004).

A promising potential application to be developed is reconstructing relative abundances from presence–absence data. For the vast majority of the global fossil record, information about relative abundances is not available and it might never even have been collected. And even if specimen counts were recorded (Moore et al., 2007; Olszewski, 2012; Tomasoyvich & Kidwell, 2011), they would not necessarily accurately reflect the abundances of living communities due to taphonomic biases (Badgeley, 1986; Damuth, 1982; Lyman, 1994) or the fragmentary nature of the remains such that they cannot be assigned to the species or even higher taxonomic levels reliably. The idea is to draw upon ecological relationships to predict what abundances could have been expected.

A recommender systems approach does not require counting frequencies of bones but instead leverages companionships of species at sites. The idea is to produce a model that can estimate the probability of occurrence for any species at any site in the dataset and then convert those estimated probabilities into relative abundances assuming that the higher the probability of occurrence is, the more abundant that species has been at that site. If the contexts of occurrence (companionships) relate to environmental patterns, then the preference scores predicted from companionships should carry abundance information. While it is clear that high environmental suitability does not always indicate high abundance and complications arise due to different kinds of rarity (Yu & Dobson, 2000) or metabolic scaling (Marquet et al., 1995), to the first approximation this relationship is true (Weber et al., 2017).

Recently, several approaches have been proposed to predict relative abundances of plants as a function of their climatic tolerances at the present day (Bradley, 2016; Couwenberghe et al., 2013; Yanez-Arenas et al., 2014). This is not directly applicable to fossil data, not least because high-resolution climatic variables are not widely available. A theoretical model for reconstructing ancestral population sizes conditioned on a phylogenetic tree was recently proposed (Manceau et al., 2020), but any practical attempt to apply this to real fossil data is still pending. In the fossil realm, analytical studies of relative abundances typically have the relative abundances available from specimen counts and are primarily concerned with estimating their confidence intervals (Buzas, 1990; Chang, 1967; Moore et al., 2007). We are not aware of any attempts to reconstruct relative abundances at large scales based solely on occurrence data of fossil species, which, among other uses, we aim at with the recommender systems approach.

2.2 | Implementation

Machine learning techniques for recommender systems have rapidly advanced over the last couple of decades (Adomavicius & Tuzhilin, 2005; Aggarwal, 2016; Ricci et al., 2011) not least due to commercial interest. Research was further catalysed by the Netflix competition, where 1 million dollars was offered and granted (Bennett & Lanning, 2007; Koren, 2009) for a film rating prediction algorithm that outperformed a contemporary movie recommender system.

Automated recommender systems are of two basic types: content based and collaborative filtering (Ricci et al., 2011; Su & Khoshgoftaar, 2009). Content-based recommender systems build personalised models that predict user preferences using product characteristics as inputs. This would correspond to the traditional

species distribution modelling, which predicts using environmental conditions as inputs.

Collaborative filtering captures and extrapolates patterns of co-occurrence. The main predictive power comes from the assumption that what occurs together is likely to occur together again without the need to explicitly describe why they occur together or define a joint morphological space. Co-occurrences link fragmented data of different users.

The predictive power comes at a price. While content-based approaches generally predict rare and common items comparably well, collaborative filtering is likely to predict rare items less accurately than common items. Collaborative filtering may be preferred in the circumstances where quantifying traits or environmental characteristics is not feasible.

Implementation of collaborative filtering offers many possibilities. Latent factor models (Gopalan et al., 2015; Hu et al., 2008; Koren et al., 2009; Ning & Karypis, 2011; Salakhutdinov & Mnih, 2008) largely dominate the collaborative filtering research for over a decade due to their simplicity and effectiveness, even though deep learning is taking over (Fu et al., 2020; Jannach et al., 2020; Liang et al., 2018; Steck, 2019), especially in commercial contexts where very large datasets are available for model training.

Latent factor models work by decomposing the user–item interaction matrix into the product of two matrices of much lower dimensionality. The principle is illustrated in Figure 1. Factorisation projects a large dataset into two smaller matrices, which summarise user preferences in a low-dimensional space. The first matrix summarises the preferences of users for items, and the second matrix summarises the affinity of items towards users. In the setting of community distribution modelling, the weights in the first matrix (**X**) represent the profiles of different sites as combinations of extracted latent types (which potentially could be interpretable, as, for example, hot or cold climate sites, or could have no interpretable meaning), while the weights of the second matrix (**Y**) represent preferences of species to those latent types of sites. Algorithmically, the main objective of this decomposition is to reconstruct the original user–item matrix not ideally, but approximately, such that the most prominent co-occurrence patterns are captured. Multiplication of the two summary matrices then produces preference estimates for all user–item (or, in the ecological context, site-species) pairs.

Exact factorisation solutions are often infeasible; thus, constructing projection matrices from data generally requires algorithmic treatment via machine learning. Tens if not hundreds of latent factor models and variants for recommender systems are available (Koren et al., 2009; Symeonidis & Zioupos, 2017), tailored for various circumstances. One criterion to consider is whether feedback about user preferences is recorded in the dataset explicitly or implicitly. Explicit feedback means that information about user preferences is available, for example, as ratings given by users to items. Implicit feedback characterises datasets where only user–item interactions are recorded, but it is not known to what extent the user liked the items. For example, a person may have purchased and watched a movie but not liked it, or they may have purchased it for someone else. Thus, the presence of a transaction does not necessarily mean high preference for the content.

Most large-scale fossil databases record only presence–absence, which in the recommender systems setting is referred to as a transaction. If a species is on the species list, it does not necessarily mean that it has been abundant or thrived in that environment, it may well have been barely surviving there. In this setting, the feedback about preferences is implicit. Explicit feedback would correspond to information about relative abundances being available. Collaborative filtering with implicit feedback can account for these uncertainties (Hu et al., 2008; Verstrepen, 2015).

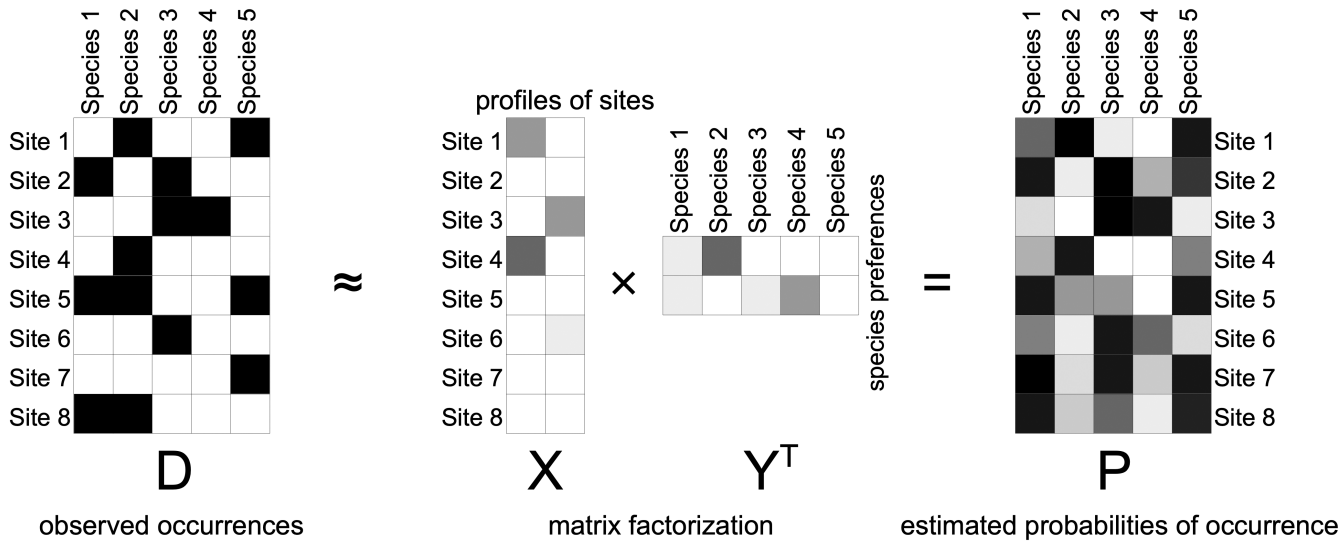

FIGURE 1 A schematic illustration of a latent factor model applied to fossil data. In the occurrence matrix, black means presence and white means absence. In the factorised matrices, shades of grey mean different weights. In the final matrix (farthest right), black means a high probability of occurrence, and white means low.

Solutions with implicit feedback typically draw on the repetitiveness of the transaction or the certainty associated with the transaction. In the fossil world, repetitiveness can be inferred, for instance, from the occurrence of species from the same genera at the same site. In addition, certainty can be quantified via qualifiers associated with species identification ('cf', 'aff' and the like). Neither certainty nor presence–absence information has to be complete; recommender systems are meant to operate on incomplete information.

2.3 | Algorithmic details

We propose to build community distribution models on a classical latent factor modelling technique for collaborative filtering with implicit feedback (Hu et al., 2008). The method works as follows.

Let $\mathbf{D}_{n\times m}$ be a matrix of observed presence or absence of species at sites. Here n is the total number of sites, and m is the total number of species in the dataset. Presence–absence in **D** most typically is binary (1 = presence, 0 = absence), but it does not have to be binary (e.g. 0.7 might encode presence with some taxonomic uncertainty).

First, presence–absence is translated into a confidence matrix $\mathbf{C}_{n\times m}$, which considers the uncertainty of absence. Hu et al. (2008) define the confidence matrix as

$$\mathbf{C} = 1 + \alpha\mathbf{D}, \tag{1}$$

where α is the parameter that accounts for asymmetry of uncertainty. This parameter roughly describes how much more a presence of a transaction is certain than an absence. Hu et al. (2008) use $\alpha = 40$ for movie recommendations. In practice, fossil data tend to have less asymmetry in uncertainties than movie recommendation data and we recommend use of $\alpha = \sim 10$ with fossil data.

The next step is matrix factorisation, which decomposes the occurrence matrix **D** into two preference matrices $\mathbf{X}_{n\times k}$ and $\mathbf{Y}_{m\times k}$ such that **X** summarises profiles of sites (as illustrated in Figure 1), and **Y** summarises preferences of species. k is a parameter specifying the dimensionality of the projection, which is conceptually similar to the number of components for any projection technique, such as principal component analysis. Our experiments suggest that $k = \sim 10$ works well for mammalian fossil data of continental scale, but it highly depends on the overall size of the data.

$$\mathbf{D} \overset{\mathbf{C}}{\rightarrow} \mathbf{X}_{n\times k} \times \mathbf{Y}^{\mathsf{T}}_{m\times k} \tag{2}$$

such that the following cost function is minimised

$$\min_{x_\star, y_\star} \sum_{u,i} c_{ui}\left(d_{ui} - X_u^{\mathsf{T}} Y_i\right)^2 + \lambda\left(\sum_u \|X\|^2 + \sum_i \|Y_i\|^2\right). \tag{3}$$

The cost function consists of two terms, the second of which is a regularisation parameter that prevents the entries of **X** and **Y** becoming too large; this is necessary to mitigate the risk of model overfitting. Here c and d are elements of matrices **C** and **D** (defined earlier), and X and Y are rows of matrices **X** and **Y**.

Factorisation is obtained algorithmically by minimising the cost function (3). First, matrices **X** and **Y** are initialised randomly and then each of their columns is repeatedly refined iterating over two equations, each time over all m species and all n sites:

$$X_u = \left(\mathbf{Y}^{\mathsf{T}}\mathbf{C}^u\mathbf{Y} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{C}^u D(u),$$
$$Y_i = \left(\mathbf{X}^{\mathsf{T}}\mathbf{C}^i\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{C}^i D^{\mathsf{T}}(i).$$

Here $\mathbf{C}^u$ is defined as an $m\times m$ diagonal matrix deriving from the confidence matrix such that $\mathbf{C}^u_i i = c_{ui}$, and $\mathbf{C}^i$ is similarly defined as an $n\times n$ diagonal matrix deriving from the confidence matrix such that $\mathbf{C}^i_{uu} = c_{ui}$. Vectors $D(u)$ and $D(i)$ are column and row vectors from the occurrence dataset **D**. $D(u)$ contains presences and absences of all the species at site u and $D(i)$ contains presences and absences of species i across all the sites. The regularisation parameter λ has the same purpose as in many statistical and machine learning optimisation procedures such as regularised regression (Hastie et al., 2009), where a small constant is added to the diagonal to avoid over-parameterisation. **I** is the identity matrix of $k\times k$ dimensions.

It is recommended to repeat the procedure for at least 10 iterations or until convergence (meaning that with further iterations the values do not change much). A source code implementing the algorithm in R along with a toy example from which Figure 1 is made is given in the online appendix.

Estimated preference scores of each species for each site can be obtained by multiplying the factor matrices

$$\mathbf{P} = \mathbf{X}\mathbf{Y}^{\mathsf{T}}. \tag{4}$$

Each element p_{ui} records a preference score of taxon i for site u. These estimates may occasionally go above one or below zero; thus, strictly speaking, they are not probabilities unless normalised (e.g. by placing a logistic function on the model output). Within this study, we do not normalise them, because it is not necessary. The interpretation can be obtained directly: the higher the score, the more likely the taxon is to have occurred at a given site. The main benefit of this raw treatment is that the estimates are obtained from a single statistical model, which makes the relative magnitude of the estimates comparable across all species and sites.

2.4 | Selecting parameters for the model

The proposed model requires four parameters: α, k, λ and the number of model fitting iterations.

α in Equation (1) controls the asymmetry of uncertainties about presence and absence. The higher the value, the more important accurate prediction of presences is considered over accurate prediction of absences, that is, higher α assigns more uncertainty

to absences. The constraint $\alpha > 0$ should be observed. $\alpha = 1$ corresponds to the assumption that absences are as equally certain as presences. Higher αs produce more of positive predictions and predictions close to zero become increasingly rare.

k in Equation (2) is the number of inner dimensions in the factor matrices. k is an integer and naturally $k > 0$. Lower k generally means more clumping of species and sites in the reconstruction of preferences for occurrence, and higher k means more individual predictions. Modelling with low k conceptually relates to the notion of chronofaunas (Bingham & Mannila, 2014; Eronen et al., 2009; Kaya et al., 2018; Olson, 1952), where sets of species make cohesive discrete units that persist over a long time and are expected to be found in their original composition at many places. The meaning of k is similar to that of the number of components in principal component analysis. In the fossil record analysis, this parameter can be interpreted as the number of distinct dietary-environmental categories.

λ in Equation (4) is the regularisation parameter, the purpose of which is to keep the elements in the two-factor matrices small. Overgrowing of the weights leads to overfitting of the observational data in the sense that the dataset is learned 'by heart' by the model instead of extracting generic patterns from it.

The number of iterations for the model fit is given in Equation (3). The model generally converged quite quickly; we found that 10 or fewer iterations were usually enough to obtain a stable model.

For exploration purposes, we tested around 200 parameter settings via a grid search in the four-dimensional model parameter space (α, k, λ and the number of model fitting iterations) using the analysis dataset. Our grid search is reported in Appendix B of this manuscript, also available as an extended abstract (Zliobaite, 2021).

For the case study, we fixed the following parameter settings: $\alpha = 10$, $k = 10$, $\lambda = 10$ and 10 iterations for model fitting. One can see from the roundedness of the values that these parameters are not fine-tuned to optimise any single quantitative criterion, but rather aimed at robust and biologically meaningful results, informed by our grid search.

We initialised matrices **X** and **Y** by drawing random values from the normal distribution with zero mean and unit variance. It took a couple of minutes to fit one model using ad hoc implementation in R suite on a commodity laptop.

For integrity and more consistent exposition in this study, we chose a single model rather than using separate models for each task of the case study or averaging over multiple random initiations. This certainly can be done, especially when such modelling is used for estimating relative abundances.

2.5 | Performance criteria for model fitting

An overarching questions is how to evaluate whether the performance of a fit model is any good. While many indirect evaluation approaches exist for recommender systems (Herlocker et al., 2004), usually the most reliable is online testing, where users are exposed to different recommender solutions at random. Since species occurrence data are almost always exclusively observational, online evaluation is not an option and we are left to evaluate the model fit based on the observational data used for modelling.

If we wanted the model to reconstruct the observational data as closely as possible, the best approach would be to set the number of internal dimensions k as high as possible and to set the regularisation parameter λ to zero. Such a model would memorise and reconstruct underlying data perfectly but it would not have generalisation or predictive power, it would simply overfit the data at hand. We want the model to extract generic patterns, not to memorise the data.

Cross-validation (Hastie et al., 2009), that would normally be used in machine learning to avoid overfitting, is not an option here since there is no trivial way to separate a testing set. For latent factor models, we can do pseudo-cross-validation, where individual occurrences are nullified at random (Ning & Karypis, 2011), and check which parameter settings best reproduce the nullified occurrences. Yet, this is not sufficient either. If we were only to maximise this leave-one-out accuracy, the optimal solution would be to predict everything as ones, that is to predict all species to occur everywhere. Clearly, this is not an informative outcome either.

Even if we manage to have a good summary that avoids overfitting, reconstructing the original observational data as closely as possible is not the only objective of the proposed community distribution modelling. Ideally, we want the model not only to reproduce observed occurrences, but also to identify the species that are most likely to be missing at sites, as well as flag potential misidentifications. Clearly, the performance criteria must include something other than just the goodness of fit. In other words, predictions must be somewhat inaccurate with respect to the training data to produce meaningful predictions. We aim at reproducing the occurrences in the original data reasonably accurately while predicting more positive occurrences than in the original data, but not too much more. While a movie recommender system could potentially keep recommending highly scored movies to the user for as long as the user keeps watching them, an informative species distribution model should recommend a finite number of species that can exist within the carrying capacity of the habitat.

There is no single quantitative criterion optimising for which would produce the most biologically meaningful model here, or in general (Warren et al., 2020). Our strategy is to construct quantitative evaluation criteria based on a subset of data points for which we have high confidence of positive occurrences and absences and allow deviations on the rest of the dataset. Given that deviations are allowed we would still want the patterns of commonness of species to be preserved, that is species that are common in the observational data should remain common, and species that are rare should remain rare.

We can assemble a subset observations for which we have high confidence from repetitive presences (e.g. occurrence of multiple species from the same genus) and absences out of a known range (e.g. a time range where species has been alive). This time information would not be used in the model itself, only in the selection of

data points for model evaluation. We can then aim at one or several conventional accuracy metrics that can assess the goodness of ranking, coming out of the model predictions on those designated true-positive and true-negative observations. The area under curve (AUC), for instance, the receiver operating characteristics curve (ROC), while not without its limitations (Hand, 2009; Peterson et al., 2008) is one popular and widely understood measure that can be used for the purpose.

Overall, we advise to consider multiple measures for model assessment, focusing on (a) accuracy of reconstructing presences and absences on a subset of observational data for which we have higher confidence coming from meta information that is not part of model fitting (time ranges, repeated occurrences); (b) accurate reproduction of positive occurrences via leave-one-out cross validation; and (c) correlation of the predicted number of species at sites with the observed number of genera, and linear correlation of the predicted number of occurrences for each genera with observed occurrences.

Finally, robustness is another desired characteristic. Model predictions should remain stable over multiple random initiations, slight perturbations of the input data or slight variations in the values of the parameters.

While one could also average over multiple random initiations for robustness of the results or even combine models with different parameter settings, the latter would produce an ensemble (Kuncheva, 2014). For clarity of exposition of the methodology, we used the same model for the case study throughout.

2.6 | Validation with present-day data

Before proceeding with the case study of fossil mammals, let us see how the proposed technique would perform on present-day data of mammal occurrences, where presences, absences and taxonomic affiliations are more certain. We aimed at a setting which would resemble our fossil case study in scope.

For this analysis, we used distribution data of large herbivorous mammals (Artiodactyla, Perissodactyla, Proboscidea and Primates) from International Union for Conservation of Nature (IUCN) Red List (IUCN, 2014). Global species occurrences were extracted from digital geospatial ranges and mapped on the grid of 50×50 km resolution, as reported in Oksanen et al. (2019)). Here a grid cell conceptually corresponds a fossil site and spatial averaging substitutes temporal averaging (Du & Behrensmeyer, 2018). We selected Africa for this analysis, the continent that hosts the richest and least deprived large mammal communities today. The continental dataset contained 8,238 grid cells for Africa. To make the size of the dataset comparable to the fossil data that used for the case study and to get rid of spatial redundancy, we randomly selected 350 grid cells with a requirement to contain at least three species. The validation dataset obtained this way included 126 species that occurred 5,437 times; on average 9.8 species occurred per cell.

We assumed that IUCN distributions of large mammals are as complete as they can be and for this exercise we considered the reported distributions to be uniformly confident in space and across species. To mimic the fossil record scenario where taxa would have been present but had not made it to the fossil record of a site we randomly nullified a certain percentage of present-day occurrences. We varied percentages and random initialisations over multiple experiments.

A desired outcome for the model is to predict high probabilities of occurrence for those nullified cases and at the same time keep predicting low probabilities of occurrence for true absences. When randomly selecting a certain percentage of presences to be nullified we at the same time randomly selected the same number of true absences, which together with the presences made validation sets. We varied the percentages of missing presences from 0.1% up to 90% and repeated each experiment 30 times with different random initialisations and random selection of missing data points.

Figure 2a shows the results. We see a clear separation between the black dots (nullified occurrences) and the white dots (true

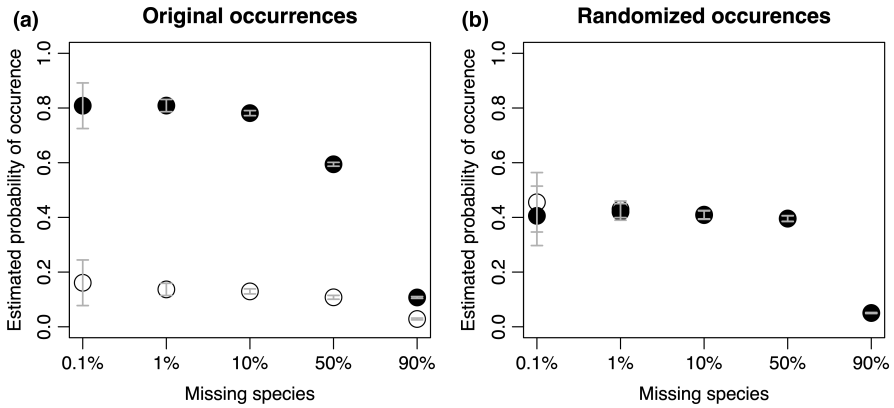

FIGURE 2 Validation on present-day occurrences of large plant eating mammals. Black dots indicate the mean estimated probability of occurrence for true presences that were, for testing purposes, masked into absences in the input data. White dots indicate the mean estimated probability of occurrences for true absences in the input data. The bars indicate the standard deviation over 30 runs where randomly selected presences were masked into absences. (a) Results on the original occurrences and (b) a sanity check with randomly shuffled occurrences (white dots are overlapping with black dots).

absences), the model performs as expected. Naturally, when the percentage of missing occurrence data is increasing, the predictions become less accurate (the black dot descends), but even at 50% missing occurrences the separation remains very clear and the standard deviation is low. The standard deviation is higher at the beginning because fewer missing values make smaller validation sets.

The model works on co-occurrence information. Thus, if there are no systematic co-occurrences (due to affinity to similar environments), the model should not work. To test this, we randomly shuffled occurrences. This experiment still contains the same number of occurrences as the original dataset but species occur and co-occur at random places, thus any real co-occurrence patterns are destroyed (for this testing purpose). We repeated the same procedure as above, nullifying a certain percentage of occurrences at a time and we did 30 runs for each missing percentage.

Figure 2b shows the same experiment with randomly shuffled occurrences. Since recommender systems draw on co-occurrence patterns, when occurrences happen at random, the predictions should not work. Indeed, we see in the figure that in this sanity test the white and the black dots are not separable.

The two experiments reassure that the methodology can recover meaningful patterns when there are such patterns and remain silent when there are no real underlying pattens.

3 | ANALYSIS OF FOSSIL MAMMALS IN MIOCENE EUROPE

Next, we demonstrate how the proposed approach can be applied to the fossil record. We show three potential application scenarios: building community trees for analysing companionships of taxa; analysing uncertainties in identification; and predicting relative abundances of taxa at sites from faunal lists.

3.1 | Fossil data and data pre-processing

Data for our case study come from the NOW (New and Old Worlds) database of fossil mammals (The NOW Community, 2021). The reported results are based on a public version downloaded on 9 March 2021.

We focus on the Middle Miocene in Europe, which gives a relatively well-resolved and curated fossil record and includes interesting faunal transitions from brachydont-dominated faunas to hypsodont-dominated faunas along with expansion of grasslands in the continent (Fortelius et al., 2014; Stromberg, 2011). The time span is arbitrarily selected to cover around 10 million years and includes a wider time span that the Middle Miocene defined stratigraphically. The dataset was filtered to include sites with the maximum age in NOW not exceeding 17.3 and minimum age not less than 7.5 million years. This captured the sites assigned to the European Land Mammal biozones from MN4 to MN12 (Hilgen et al., 2012), as well as individually dated sites falling within this age range. We discarded sites that had an age range of more than 2.3 million years, which is the maximum time span among the MN units.

We selected European sites by country (Switzerland, Spain, Greece, Germany, Italy, France, Turkey, Cyprus, Bulgaria, Russia, Georgia, Ukraine, Austria, Portugal, Belgium, Romania, Moldova, Azerbaijan, Armenia, United Kingdom, Hungary, Poland, Serbia, Slovakia, Netherlands, Czech Republic, Croatia, Malta, North Macedonia, Norway, Serbia, Montenegro, Sweden, Finland and Belarus) and excluded everything that is to the east of 45° longitude, which is roughly where the Volga River in Russia flows.

We restricted our analysis to large herbivores by their affiliation to Artiodactyla, Perissodactyla, Proboscidea, Primates or Hyracoidea. We formed the faunal lists at sites at the genus level. This treatment makes no difference from an algorithmic perspective, but it is more prudent from a palaeontological perspective. Genera have been argued to be more robust than species for palaeoecological analyses (Eronen et al., 2011), and many previous large-scale studies on NOW data used genera as the basis (Eronen et al., 2009; Fortelius et al., 2016; Jernvall & Fortelius, 2002; Kaya et al., 2018; Zliobaite et al., 2017).

The NOW database lists synonyms at the species level; thus, when aggregating at the genus level, synonyms, if any, are not taken into account. We discarded entries for which the genus was indetermined ('indet.') or unspecified ('gen.'). We did not account for synonyms, except for one case which was necessary to facilitate analysis of relative abundances at Pasalar, Turkey: we replaced two occurrences of *Procoelodonta* with *Begertherium* by script.

We discarded sites that had fewer than three genera and then discarded genera that occurred in fewer than three sites. This left us with a dataset covering 351 sites with 104 genera. The dataset included 2,746 observations of genus-site occurrences, on average 7.8 genera per site.

We accounted for the uncertainty of occurrences in the following way. If no additional information was available, for any genus listed at a site we arbitrarily assigned an observed probability of occurrence of 0.9. If several species of the same genus were reported to occur at the same site, we set the probability of occurrence of that genus to 1.0. For any genus that was listed but the species was undetermined, we assigned a probability of 0.7. The same applied to any genus that had uncertain species validity or was listed as informal species in the database. If, in the database, genus attribution of species, family attribution of genus or taxonomic validity was uncertain, then we set the probability of occurrence of that genus at that site to 0.5.

3.2 | Analysis workflow

Fitting the recommender systems model gave us a matrix of preference scores of genera for sites, which we used for three types of analysis.

Our first analysis aimed at building community trees, identifying companionships of genera and analysing which genera often

occur together. This task is remotely related to the concept of chronofaunas (Eronen et al., 2009; Kaya et al., 2018; Olson, 1952). Chronofaunas are designated assemblages of taxa that all occurred at the same time at some classic site. Chronofaunas do not assess tendencies to occur together and may include some taxa that generally are not likely to occur together but which happened to occur together at that particular place. Taking the concept of chronofauna further, rather than arbitrarily forcing taxa into discrete units of co-occurrence, we assessed the tendency of genera to occur together and made a community tree out of that, where closeness of taxa on the branches of that tree signals that they are more likely to occur in each other's companionship.

For making the tree, we used internal components of the model rather than its output scores. Recall from Equation (2) that the model is $\mathbf{D} = \mathbf{X}\mathbf{Y}^{\mathbf{T}}$, where **D** is the matrix of preference scores, **X** and **Y** are summary profiles of sites and genera, respectively. Our analysis used matrix **Y**, which sometimes can be seen as a by-product of matrix factorisation. We ran hierarchical clustering on **Y** with the Euclidean distance and Ward's linkage (Ward, 1963), which minimises the variance when joining branches. We chose this linkage for its robustness.

The resulting community tree can be viewed as a 'phylogeny' of companionships (in contrast to regular phylogeny by ancestry). One could do a similar clustering exercise on raw occurrence data, but then the clustering reflects the realised occurrences rather than their (inferred) preferences. The former would be much more driven by the absolute number of occurrences, and the latter is more driven by the context of occurrences.

Our second analysis primarily served as an additional model validation task. We used the model output scores to analyse undetermined occurrences of cervids. We selected all cases marked as Cervidae indet. at the sites within the analysis dataset and investigated which of the cervids had the highest propensity scores for those sites. This is by no means meant to replace taxonomic identification of those remains, but it may give interesting insights into patterns of occurrence. We chose Cervidae because of a relatively common case in the Middle Miocene European record in which a genus was identified as either *Euprox* or *Heteroprox*, which are hard to distinguish morphologically. Thus, in addition to our analysis of undetermined cervids, we specifically looked at the four sites where those undetermined cervids were annotated as either *Euprox* or *Heteroprox*, but it was undetermined which one. We analysed whether the model could make a distinction between those two.

Our final analysis aimed to predict the relative abundances from faunal lists. We used the model scores for ranking genera at sites by their estimated propensity to occur. We assumed that the higher the propensity to occur, the higher the relative abundance. This is quite a strong assumption, which we have to make in order to be able to link propensity to occur to relative abundances. By and large this holds in community ecology, even though in reality, abundance–distribution relationships can be much more complex (Weber et al., 2017).

We mapped the ranking of preferences of taxa from the recommender systems model versus relative abundances, calibrating it

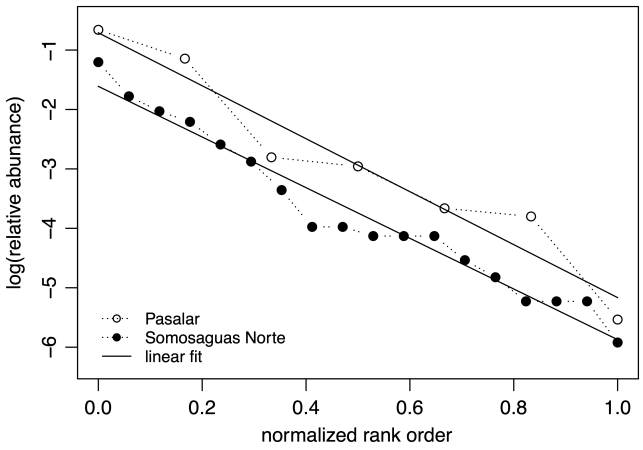

FIGURE 3 Distribution of observed relative abundances at Somosaguas Norte in Spain (MN5) and Pasalar in Turkey (MN6). Normalised rank order means that the relative abundances are ranked from the highest to the lowest and rank orders are normalised to fall between 0 and 1, where 0 represents the most abundant taxa and 1 represents the least.

based on two fossil sites for which specimen counts were reported in the literature: Somosaguas Norte in Spain (MN5) (Domingo et al., 2017) and Pasalar in Turkey (MN6) (Andrews & Ersoy, 1990). Figure 3 depicts the distribution of observed relative abundances at these two sites. We see that relative abundances follow a surprisingly regular geometric relationship. This relationship is consistent across both sites as indicated by a tight fit of the trend line with almost identical slopes (−4.46 for Pasalar, and −4.26 for Somosaguas Norte). The intercepts differ because the total number of distinct taxa is very different across both sites. Somosaguas Norte has 7 and Pasalar has 18 distinct large mammal species reported. The relationship is delightfully consistent with the classical model of relative species abundances proposed almost a century ago (Motomura, 1932).

This information about relative abundances is only available for 2 out of 351 analysis sites. Those relative abundances have not been used in any way for fitting a recommender systems model and will serve us for assessing predictions of relative abundances. Despite the regularity of the relationship in Figure 3, we must keep in mind that these relative abundances do not necessarily represent the ground truth that we want to predict, that is, relative abundances within faunal communities that lived. The relative abundances in Figure 3 are filtered by taphonomic processes. Thus, we expect some match between the recommender systems model predictions and the observed relative abundances, but we do not aim for an ideal match.

In addition, there is the challenge of matching taxon names. Somosaguas Norte has all the genera matching the taxonomic names in the NOW database. *Retroporcus* of Somosaguas Norte was excluded from our analysis by the filter requiring a genus to occur in at least three sites (it had only one occurrence). The published Pasalar species count has some mismatches of taxonomic names with those reported in the NOW database. The

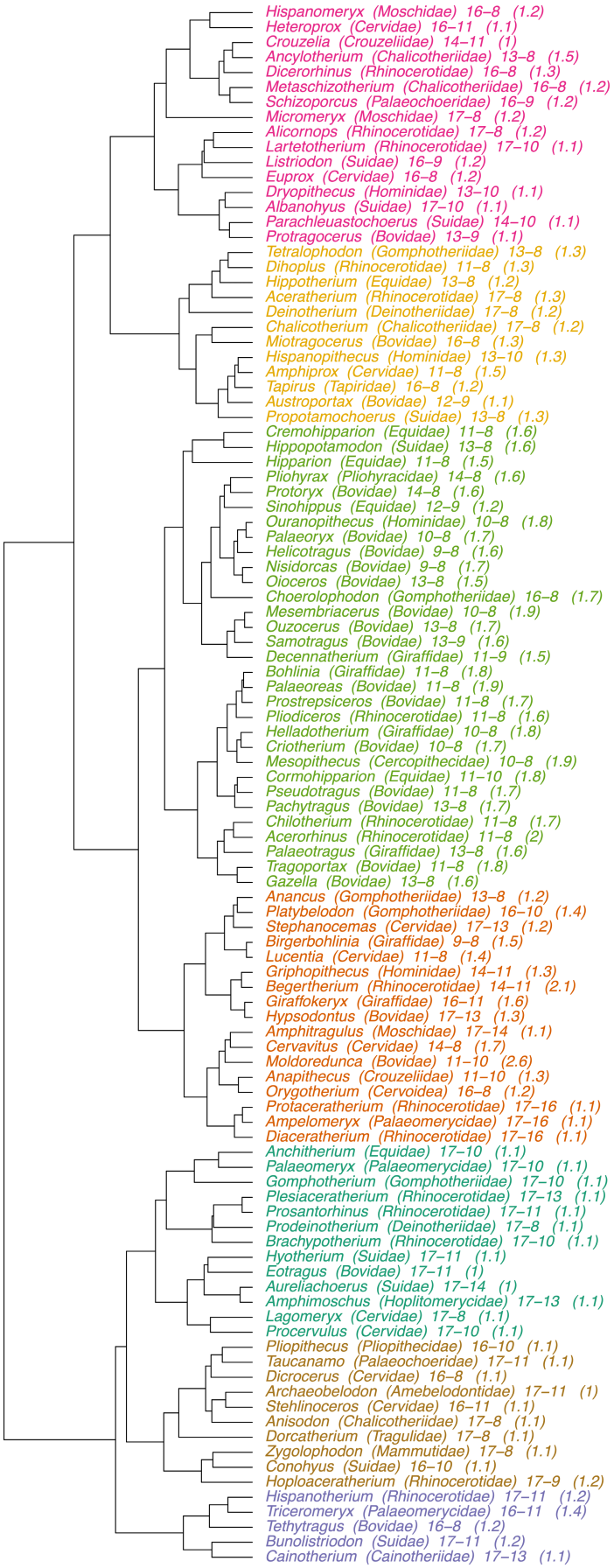

FIGURE 4 Companionships of genera according to their propensity to occur. The numbers following the family indicate approximate ages of genera in millions of years, and the numbers in brackets indicate the mean ordinated hypsodonty of their occurrence sites. The mean ordinated hypsodonty excludes the genera for which the context is computed. Colour coding is for readability; it arbitrarily distinguishes the seven clusters.

TABLE 1 Propensity to occur for cervids at sites where Cervidae indet. is listed. Grey shading indicates overlapping time ranges. The highest scores for each site are indicated in bold. Underlined are existing occurrences of cervids at those sites in addition to Cervidae indet. Stars indicate the sites where Cervidae indet. has additional information listed. HYP = ordinated hypsodonty

		Maximum age, ma	Minimum age, ma	Mean HYP	*Stephanocernas*	*Procervulus*
Maximum age, Ma					17.2	17.2
Minimum age, Ma					12.8	9.9
HYP context					1.2	1.1
Baggersee Freudenegg 2	Germany	17.2	16.4	1	0.05	0.33
Baggersee Freudenegg 3	Germany	17.2	16.4	1	0.10	0.24
Torralba 1	Spain	17.2	16.4	1	0.03	**0.48**
Armantes 1	Spain	17.2	16.4	1.2	0.10	**0.58**
Munebrega 1	Spain	17.2	16.4	1.3	0.13	**0.58**
Enghausen*	Germany	16.4	14.2	1	0.10	0.25
Oggenhausen 1*	Germany	16.4	14.2	1	0.08	0.53
Ziemetshausen 1b*	Germany	16.4	14.2	1	0.16	0.42
Hullistein	Switzerland	16.4	14.2	1	0.09	**0.34**
Armantes 3	Spain	16.4	14.2	1	0.11	**0.47**
La Retama	Spain	16.4	14.2	1.3	0.13	**0.63**
Murero	Spain	14.2	12.8	1	0.04	0.09
Gisseltshausen 1a*	Germany	14.2	12.8	1	0.01	0.42
Petersbuch 6	Germany	12.8	11.2	1	0.01	0.10
Saint-Gaudens (Valentine)	France	12.8	11.2	1	0.12	0.33
Hammerschmiede**	Germany	12.1	11.5	1	0.11	0.07
Creu Conill 20	Spain	11.2	11.1	1.4	−0.04	−0.20
El Lugarejo	Spain	11.2	9.9	2	−0.09	0.20
Draxeni***	Moldova	11.2	9.9	1.7	−0.05	0.16
Sinap 91	Turkey	11.1	10.0	2	0	**0.28**
Montredon	France	9.9	8.9	1.6	0	0.05
Sinap 1	Turkey	9.7	9.6	1.5	0.03	**0.31**
Vivero de Pinos	Spain	8.9	7.6	1.7	−0.01	0.02

Notes: **Euprox/Heteroprox*; **Muntiacini; ***Lagomerycinae.

recommender systems model follows the current faunal list from NOW.

For the purpose of the present pilot study, we map reference scores to relative abundances at site *j* in a simplified way as follows:

$$\pi_{ij} = \frac{(p_{ij} - 0.5)}{\sum_{d_{ij}>0} p_{ij}}, \quad (5)$$

where p_{ij} is the preference score for species *i* to occur at site *j*, coming from the model; d_{ij} is the presence–absence matrix, where $d_{ij} > 0$ means that we only sum taxa that are reported to be present at site *j*. The subtraction of 0.5 from the probability score is an arbitrary cut-off implying that preference scores below 0.5 signal absence. Although, in principle, preference scores would allow the inclusion of absent species which had high estimated probability of occurrence, in this pilot study we only analyse the relative abundances of present species.

3.3 | Companionships of species

Figure 4 shows the resulting community tree of companionships. Just like phylogenetic trees depict taxa that have recent common ancestors close to each other, community trees are meant to depict taxa that occur in similar contexts and similar environments close to each other.

The numbers on the branches give age of the genera as well as the mean ordinated hypsodonty of their sites as a proxy for their environmental conditions. Neither the age nor the hypsodonty scores were used in clustering; the resulting hierarchical tree is from their site preference scores only.

We can see that the grouping of species matches in age, but not too closely, which gives an interesting perspective; perhaps some of those genera could have occurred together for longer if one of them had not gone extinct or been replaced. Curiously, we can see that *Heteroprox* and *Euprox*, which are often

Lagomeryx	*Heteroprox*	*Stehlinoceros*	*Dicrocerus*	*Euprox*	*Cervavitus*	*Amphiprox*	*Lucentia*
17.2	16.4	16.4	16.4	16.4	13.6	11.2	10.6
7.6	11.2	11.1	8.0	7.6	7.6	7.6	7.6
1.1	1.1	1.1	1.1	1.2	1.7	1.5	1.4
0.42	0.35	0.08	0.14	0.04	−0.06	−0.06	−0.01
0.40	0.23	0.16	0.26	−0.06	−0.07	0.02	0.07
0.44	0.07	−0.05	0	0.25	0.01	0.03	−0.06
0.41	0.22	0.09	0.15	0.44	−0.10	0.04	0.07
0.50	0.06	0.04	0.13	0.05	−0.06	−0.03	0.02
0.31	0.16	0.02	0.02	−0.06	0.01	0.08	0.10
0.58	0.15	0.05	0.18	−0.08	0.07	0.09	0.14
0.76	0.15	0.46	**0.86**	0.21	0.02	0.11	0.13
0.33	0.20	0.10	0.15	0.01	−0.10	−0.01	0.06
0.44	0.08	0.06	0.17	0.11	−0.03	0.06	0.11
0.32	−0.01	−0.13	−0.24	0.11	0.04	−0.01	−0.03
0.12	0.44	0.19	0.13	**0.51**	−0.06	0	−0.08
0.68	0.36	0.08	0.32	0.30	0.12	0.02	−0.04
0.39	0.32	0.01	0.21	0.23	0.09	0.06	0.05
0.21	0.32	0.49	0.47	**0.62**	0.04	0.20	0.12
−0.19	0.10	0.33	0.23	**0.78**	−0.07	0.27	0.20
−0.13	0	0.11	0.16	**0.78**	0.09	0.38	0.17
−0.05	0.25	−0.04	−0.10	**0.24**	−0.07	0	0.07
0.23	−0.04	−0.08	0.04	0.24	0.18	**0.29**	0.19
0.05	0.09	0.05	−0.03	0.24	0.22	0	−0.10
0.14	0.18	0.10	0.28	**0.58**	0.18	0.36	0.34
0.17	0.02	0.06	0.04	0.27	0.30	0.04	−0.10
0.26	0	0.12	**0.41**	0.35	0.02	0.25	0.21

indistinguishable morphologically, are relatively far apart in terms of their companionships.

We also see from the tree that the closest companionships are not formed within the same families, which might relate to the principle of competitive exclusion (Hardin, 1960), which suggests that closely related taxa are less likely to coexist.

In the figure, the companionships of genera appear to relate to their mean hypsodonty context much more closely than to their age. The mean ordinated hypsodonty relates to the aridity of the environment (Fortelius et al., 2002)—the higher the number, the more harsh the environment is. The mean ordinated hypsodonty calculation here excludes self; thus, the scores are relative to each genera and are not the same for each genera at the same site. Thus, the pattern that we see in the figure suggests that the companionships may not necessarily be about interacting with each other but rather about being in similar environments.

The colour coding arbitrarily indicates the seven largest clusters, which roughly can be thought of as another way to computationally identify chronofauna (Bingham & Mannila, 2014). Each chronofauna, represented by a different colour, does not directly correspond to any particular site but rather gives somewhat cohesive virtual faunal assemblages. We see that those virtual chronofaunas are much more compact in the Early and Middle Miocene than in the Late Miocene (green). The latter shows tight companionships of fauna with somewhat more varying hypsodonty contexts than in earlier times. That could suggest that Early Miocene faunas in Europe were more distinct and that genera of Late Miocene faunas, after grassland expansion (Fortelius et al., 2014; Stromberg, 2011), became more intermixed in their occurrences and more interchangeable.

3.4 | Possible and improbable genera at sites

Table 1 lists sites where undetermined Cervidae indet. is present and gives the model score for each potential genus of Cervidae to

occur at those sites. The higher the numerical value, the more likely that genus is to be present at that site. Since the scores come from a single model, they are comparable across genera and across sites. The highest score among cervid genera for each site is indicated in bold; those are the most likely candidates coming from the model to stand behind Cervidae indet. in each case. These predictions are by no means to be used for curating the species lists, but they can give insights into occurrence patterns of selected taxonomic groups.

3.5 | Predicting relative abundances from presence–absence

Table 2 reports observed specimen counts, preference scores coming from the recommender systems model and relative abundances corresponding to both. We can see that the match is not very close, but high-level patterns are notable. For example, *Gomphotherium* and *Anchitherium* are clearly much more abundant in the estimates coming from the model than they are in observed specimen counts. In Pasalar, the predictive patterns are less clear. In some cases, such as *Listriodon*, the observed relative abundances from specimen counts almost match the predictions if we sum *Listriodon* and *Bunolistriodon* in the predictions. But there are extreme mismatches as well, such as, for instance, *Micromeryx* or *Gomphotherium*, which are predicted to be very common at that site but are not so common according to the specimen counts.

Overall, relative abundance predictions for Pasalar are quite flat, with many of them around 5%–8%, which is a result of a long faunal list pulling down high runaway predictions. If the main target of this modelling was accurate reconstruction of relative abundances, one could have added nonlinear components to the model outputs to force more extreme predictions. We left future research to aim for a generic model which would be robust for constructing community trees.

4 | DISCUSSION

We proposed a new direction for community distribution modelling, tailored applicable for analysing fossil communities. We demonstrated the potential of the proposed analytical framework with a case study of large herbivores in Middle Miocene sites in Europe, yet many possibilities remain both for methodological development and its application.

First of all, we did not incorporate time information into the recommender systems model in any way. In principle, one could constrain the model to predict more aggressively at times when the taxa are known to have been alive and predict more conservatively outside those times. Yet even now, when time information was not used at all, predictions fell within their times well enough, as can be seen for instance from the community tree in Figure 2, where contemporary genera tend to be close together on the tree.

In addition to time constraints, other explicit information can be included in the modelling; for example, morphological traits of taxa, such as teeth, could be incorporated into community distribution modelling. This way, companionship of taxa can be represented not only in terms of co-occurrence patterns featuring other taxa, but also contexts and distributions of traits across communities. In the community tree in Figure 4, we plotted taxa along with the mean hypsodonty within their companionships. Even though this information was not explicitly part of the recommender systems model, we could see that those taxa that occur in similar trait contexts appear close together in the community tree. Hybrid recommender system techniques (Burke, 2002) could be tailored for this purpose.

The synonymity challenge is another aspect that we have not addressed explicitly in this study. Analysis of companionships of the proposed kind has the potential to highlight potential synonyms. Even more importantly, it has the potential to disambiguate possible synonymity cases, as we demonstrated with the *Euprox*/*Heteroprox* case. While morphologically those two genera are sometimes indistinguishable, the community tree showed that their companionships are quite distinct and that they tend to occur in different ecological contexts.

One could take the analysis of companionships even further. The proposed recommender systems approach allows the extraction of companionship weights for each individual taxa and those weights can be positive or negative. Although co-occurrence in the fossil record does not immediately imply competitive interactions, analysing those weights might provide a quantitative perspective on the magnitude and directionality of competitive interactions within fossil faunal communities.

While co-occurrence does not necessarily indicate biotic interactions (Blanchet et al., 2020), a community distribution model could potentially be used to infer and analyse competitive interactions. One can extract regression-like equations from the recommender model, where probability of occurrence of a species at a site is described as a function of other species. After controlling for the type of environment, one could argue that positive weights indicate competitive interactions, negative weights indicate competitive exclusion and zero weights indicate no interactions even if species co-occur. This is not part of the present study, it remains for future work.

Another possible extension could be towards using the propensity scores for estimating and comparing the sizes of geographical ranges of taxa.

An important highlight of the proposed methodology is that corrections for sampling intensity (Chao & Jost, 2012; Connolly & Miller, 2001; Raup, 1975), often administered to fossil data, are not needed here. The original recommender systems approach is tailored to work with incomplete profiles, this way there is room to recommend and accommodate other items than those that are there already. There is a signal in incompleteness. Our experiments with

TABLE 2 Reported specimen counts in Somosaguas Norte in Spain (MN5) and Pasalar in Turkey (MN6), along with our estimates for relative abundances produced from species lists via a latent factor model

Genus	Family	Number of specimens	Relative abundance	Predicted preference score	Predicted relative abundance
Somosaguas Norte, Spain (MN5)					
Gomphotherium	Gomphotheriidae	786	51%	0.88	31%
Anchitherium	Equidae	484	31%	1.02	33%
Prosantorhinus	Rhinocerotidae	92	6%	0.40	<1%
Retroporcus	Suidae	79	—	—	—
Tethytragus	Bovidae	39	3%	0.65	10%
Micromeryx	Moschidae	34	2%	0.75	16%
Heteroprox	Cervidae	6	<1%	0.66	10%
Pasalar, Turkey (MN6)					
Caprotragoides	Bovidae	112	30%	—	—
Tethytragus	Bovidae	—	—	0.98	8%
Listriodon	Suidae	63	17%	1.08	9%
Bunolistriodon	Suidae	—	—	0.95	6%
Sivapithecus	Hominidae	49	13%	—	—
Griphopithecus	Hominidae	—	—	0.55	2%
Hypsodontus	Bovidae	41	11%	0.91	6%
Giraffokeryx	Giraffidae	28	8%	0.83	5%
Anchitherium	Equidae	21	6%	1.06	8%
Anisodon	Chalicotheriidae	—	—	0.80	4%
Pliohyrax	Pliohyracidae	—	—	0.54	1%
Begertherium	Rhinocerotidae	13	3%	0.53	<1%
Gomphotherium	Gomphotheriidae	7	2%	0.99	7%
Stephanocemas	Cervidae	7	1%	0.33	<1%
Dorcatherium	Tragulidae	6	2%	0.93	6%
Micromeryx	Moschidae	6	2%	1.05	8%
Conohyus	Suidae	6	2%	0.85	5%
Brachypotherium	Rhinocerotidae	4	1%	0.89	6%
Taucanamo	Palaeochoeridae	3	1%	0.81	4%
Aceratherium	Rhinocerotidae	2	<1%	—	—
Hoploaceratherium	Rhinocerotidae	—	—	1.16	9%
Deinotherium	Deinotheriidae	2	1%	0.84	5%
Palaeomeryx	Palaeomerycidae	2	1%	0.82	5%
Hispanomeryx	Moschidae	1	<1%	0.51	<1%

present-day data reported in this study show that even with around 50% of randomly missing data we can recover missing absences and distinguish them from true absences well. This argument relies on the assumption that species are missing independently from each other, it becomes more tricky if incompleteness is correlated. The same holds for rarefaction or methods alike.

5 | CONCLUSIONS

We presented a new perspective on fossil community distribution modelling. The proposed approach leverages recommender systems of machine learning for modelling community distribution via their companionships of occurrence. Rather than treating chronofaunas as monolithic units, this approach models companionships of each taxa individually yet generalises over them in a single model. One model makes preference scores directly comparable across taxa and sites. This perspective allows the analysis of fossil communities and the relationships between their ecological contexts at high resolution as well as at large scales.

Our case study of mammalian fossil faunas of Miocene Europe showed a proof of concept how, even coarsely, relative abundances as well as ecological contexts can be reconstructed from faunal lists at sites.

One of the main applications of the proposed approach is construction of a community tree, which can be viewed as a counterpart to a phylogenetic tree. Instead of depicting ancestry relationships, a community tree depicts ecological relationships between taxa.

While many possibilities and challenges in extending the recommender system approach to fossil faunas still await, we hope our treatment will encourage new research into a faunal community perspective on fossil species distribution modelling.

ACKNOWLEDGEMENT

This research was supported by the Academy of Finland (grants 314803 and 341623). Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST

There are no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets along with analysis scripts are available on GitHub: https://github.com/zliobaite/fossilrec, https://doi.org/10.5281/zenodo.6576483 Zliobaite (2022).

ORCID

Indrė Žliobaitė https://orcid.org/0000-0003-2427-5407

REFERENCES

Adomavicius, G., & Tuzhilin, A.(2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749.

Aggarwal, C.(2016). *Recommender systems: The textbook*. Springer International Publishing.

Alroy, J.(2000). New methods for quantifying macroevolutionary patterns and processes. *Paleobiology*, *26*(4), 707–733.

Andrews, P., & Ersoy, A.(1990). Taphonomy of the miocene bone accumulations at Pasalar, Turkey. *Journal of Human Evolution*, *19*, 379–396.

Badgeley, C.(1986). Counting individuals in mammalian fossil assemblages from fluvialenvironments. *PALAIOS*, *1*(3), 328–338.

Beasley, E., & Maher, S.(2019). Small mammal community composition varies among Ozark glades. *Journal of Mammalogy*, *100*(6), 1774–1782.

Bennett, J., & Lanning, S.(2007). The netflix prize. *Proceedings of KDD Cup and Workshop*, *2007*, 35–52.

Bingham, E., & Mannila, H.(2014). Towards computational techniques for identifying candidate chronofaunas. *Annales Zoologici Fennici*, *51*(1), 43–48.

Blanchet, F. G., Cazelles, K., & Gravel, D.(2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063.

Bradley, B. A.(2016). Predicting abundance with presence-only models. *Landscape Ecology*, *31*, 19–30.

Burke, R.(2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331–370.

Buzas, M.(1990). Another look at confidence limits for species proportions. *Journal of Paleontology*, *64*, 842–843.

Chang, Y. M.(1967). Accuracy of fossil percentage estimation. *Journal of Paleontology*, *41*(2), 500–502.

Chao, A., Hsieh, T., Chazdon, R., Colwell, R., & Gotelli, N.(2015). Unveiling the species-rank abundance distribution by generalizing the good-turing sample coverage theory. *Ecology*, *96*(5), 1189–1201.

Chao, A., & Jost, L.(2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology*, *93*, 2533–2547.

Connolly, S., & Miller, A.(2001). Joint estimation of sampling and turnover rates from fossil databases: Capture-mark-recapture methods revisited. *Paleobiology*, *27*(4), 751–767.

Couwenberghe, R. V., Collet, C., Pierrat, J. C., Verheyen, K., & Gegout, J. C.(2013). Can species distribution models be used to describe plant abundance patterns?*Ecography*, *36*, 665–674.

Damuth, J.(1982). Analysis of the preservation of community structure in assemblages of fossil mammals. *Paleobiology*, *8*(4), 434–446.

Darwin, C.(1859). *On the origin of species by means of natural selection, or the preservation of Favoured races in the struggle for life*. John Murray.

Domingo, M., Martin-Perea, D., Domingo, L., Cantero, E., Cantalapiedra, J., Garcia-Yelo, B., Gomez-Cano, A., Alcalde, G., Fesharaki, O., & Hernandez-Fernandez, M.(2017). Taphonomy of mammalian fossil bones from the debris-flow deposits of Somosaguas-North (Middle Miocene, Madrid Basin, Spain). *Palaeogeography, Palaeoclimatology, Palaeoecology*, *465*, 103–121.

Du, A., & Behrensmeyer, A. K.(2018). Spatial, temporal and taxonomic scaling of richness in an eastern african large mammal community. *Global Ecology and Biogeography*, *27*(9), 1031–1042.

Dunstan, P., Foster, S., & Darnell, R.(2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, *222*(4), 955–963.

Elith, J., & Leathwick, J.(2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697.

Eronen, J., Ataabadi, M. M., Micheels, A., Karme, A., Bernor, R., & Fortelius, M.(2009). Distribution history and climatic controls of the late miocene pikermian chronofauna. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(29), 11867–11871.

Eronen, J., Evans, A., Fortelius, M., & Jernvall, J.(2011). Genera are often better than species for detecting evolutionary change in the fossil record: A reply to Salesa et al. *Evolution*, *65*(5), 1514–1516.

Foote, M.(2000). Origination and extinction components of taxonomic diversity: General problems. *Paleobiology*, *26*(S4), 74–102.

Foote, M.(2016). On the measurement of occupancy in ecology and paleontology. *Paleobiology*, *42*(4), 707–729.

Fortelius , M. , Eronen , J. , Jernvall , J. , Liu , L. , Pushkina , D. , Rinne , J. , Tesakov , A. , Vislobokova , I. , Zhang , Z., & Zhou , L. (2002). Fossil mammals resolve regional patterns of Eurasian climate change over 20 million years. *Evolutionary Ecology Research*, *4*, 1005–1016.

Fortelius, M., Eronen, J., Kaya, F., Tang, H., Raia, P., & Puolamaki, K.(2014). Evolution of Neogene mammals in Eurasia: Environmental forcing and biotic interactions. *Annual Review of Earth and Planetary Sciences*, *42*(1), 579–604.

Fortelius, M., Zliobaite, I., Kaya, F., Bibi, F., Bobe, R., Leakey, L., Leakey, M., Patterson, D., Rannikko, J., & Werdelin, L.(2016). An ecometric analysis of the fossil mammal record of the turkana basin. *Philosophical Transactions of the Royal Society: Biological Sciences*, *371*(1698), 20150232.

Z. Fu, Gao, H., Guo, W., Jha, S., Jia, J., Liu, X., Long, B., Shi, J., Wang, S., & Zhou, M.(2020). Deep learning for search and recommender systems in practice. In *Proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, KDD'20 (pp. 3515–3516).

Gopalan, P., Hofman, J., & Blei, D.. (2015). Scalable recommendation with hierarchical poisson factorization. In *Proceedings of the thirty-first conference on uncertainty in artificial intelligence* (pp. 326–335).

Hand, D.(2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hardin, G.(1960). The competitive exclusion principle. *Science*, *131*(3409), 1292–1297.

Hastie, T., Tibshirani, R., & Friedman, J.(2009). *The elements of statistical learning: Data mining, inference, and prediction*(2nd ed.). Springer.

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J.(2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53.

Hilgen, F., Lourens, L., & vanDam, J.(2012). The neogene period. In F. Gradstein, J. Ogg, M. Schmitz, & G. Ogg(Eds.), *The geologic time scale 2012*(pp. 923–978). Elsevier.

Y. Hu, Koren, Y., & Volinsky, C.. (2008). Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE international conference on data mining (ICDM 2008)*.

Hui, F., Taskinen, S., Pledger, S., Foster, S., & Warton, D.(2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*(4), 399–411.

IUCN. (2014). *The IUCN Red List of threatened species*. Retrieved from http://www.iucnredlist.org

Jannach, D., deSouza, P., & Oldridge, E.(2020). Why are deep learning models not consistently winning recommender systems competitions yet? A position paper. *Proceedings of the Recommender Systems Challenge*, *2020*, 44–49.

Jernvall, J., & Fortelius, M.(2002). Common mammals drive the evolutionary increase of hypsodonty in the neogene. *Nature*, *417*, 538–540.

Kaya, F., Bibi, F., Zliobaite, I., Eronen, J., Tang, H., & Fortelius, M.(2018). The rise and fall of the old world savannah fauna and the origins of the african savannah biome. *Nature Ecology and Evolution*, *2*, 241–246.

Koren, Y.(2009). The bellkor solution to the netflix grand prize. *Netflix Prize Documentation*, *81*(2009), 1–10.

Koren, Y., Bell, R., & Volinsky, C.(2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.

Kuncheva, L.(2014). *Combining pattern classifiers: Methods and algorithms*(2nd ed.). John Wiley & Sons.

Liang, D., Krishnan, R., Hoffman, M., & Jebara, T.(2018). Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, WWW'18 (pp. 689–698).

Lyman, R.(1994). Relative abundances of skeletal specimens and taphonomic analysis of vertebrate remains. *PALAIOS*, *9*(3), 288–298.

Manceau, M., Gupta, A., Vaughan, T., & Stadler, T.(2020). The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data. *Journal of Theoretical Biology*, *509*, 110400.

Marquet, P. A., Navarrete, S. A., & Castilla, J. C.(1995). Body size, population density, and the energetic equivalence rule. *Journal of Animal Ecology*, *64*(3), 325–332.

May, R.(1988). How many species are there on earth?*Science*, *241*(4872), 1441–1449.

Moore, J., Norman, D., & Upchurch, P.(2007). Assessing relative abundances in fossil assemblages. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *253*, 317–322.

Motomura , I. (1932). A statistical treatment of associations . *Japanese Journal of Zoology*, *4*, 166–171.

Myers, C., Stigall, A., & Lieberman, B.(2015). PaleoENM: Applying ecological niche modeling to the fossil record. *Paleobiology*, *41*(2), 226–244.

Ning, X., & Karypis, G.(2011). Slim: Sparse linear methods for top-n recommender systems. In *IEEE international conference on data mining*(pp. 497–506). ICDM.

Oksanen, O., Zliobaite, I., Saarinen, J., Lawing, A., & Fortelius, M.(2019). A humboldtian approach to life and climate of the geological past: Estimating palaeotemperature from dental traits of mammalian communities. *Journal of Biogeography*, *46*(8), 1760–1776.

Olson, E.(1952). The evolution of a permian vertebrate chronofauna. *Evolution*, *6*(2), 181–196.

Olszewski, T.(2012). Remembrance of things past: Modeling the relationship between species? Abundances in living communities and death assemblages. *Biology Letters*, *8*, 131–134.

Peterson, A. T., Papes, M., & Soberon, J.(2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modeling*, *213*, 63–72.

Pollock, L., Tingley, R., Morris, W., Golding, N., O'Hara, R., Parris, K., Vesk, P., & McCarthy, M.(2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution*, *5*(5), 397–406.

Raup, D.(1975). Taxonomic diversity estimation using rarefaction. *Paleobiology*, *1*(4), 333–342.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P.(2011). *Recommender systems handbook*. Springer.

Saarinen, J., Oksanen, O., Zliobaite, I., Fortelius, M., DeMiguel, D., Azanza, B., Bocherens, H., Luzon, C., Solano-García, J., Yravedra, J., Courtenay, L. A., Blain, H.-A., Sanchez-Bandera, C., Serrano-Ramos, A., Rodriguez-Alba, J. J., Viranta, S., Barsky, D., Tallavaara, M., Oms, O., … Jimenez-Arenas, J. M.(2021). Pliocene to middle pleistocene climate history in the Guadix-Baza basin, and the environmental conditions of early homo dispersal in europe. *Quaternary Science Reviews*, *268*, 107132.

Salakhutdinov , R. , & Mnih , A. (2008). Probabilistic matrix factorization . In *Advances in neural information processing systems* (pp. 1257 – 1264). Mitpress.

Steck , H. (2019). Embarrassingly shallow autoencoders for sparse data . In *Proceedings of the 2019 world wide web conference*, WWW'19 (pp. 3251 – 3257).

Stromberg, C.(2011). Evolution of grasses and grassland ecosystems. *Annual Review of Earth and Planetary Sciences*, *39*(1), 517–544.

Su, X., & Khoshgoftaar, T.(2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*, 1–19.

Symeonidis, P., & Zioupos, A.(2017). *Matrix and tensor factorization techniques for recommender systems*. Springer.

The NOW. (2021). Community. New and old worlds database of fossil mammals (now). Licensed under CC BY 4.0. Retrieved from http://www.helsinki.fi/science/now/

Tikhonov, G., Opedal, A., Abrego, N., Lehikoinen, A., Jonge, M. D., Oksanen, J., & Ovaskainen, O.(2020). Joint species distribution modelling with the r-package Hmsc. *Methods in Ecology and Evolution*, *11*(3), 442–447.

Tomasoyvich, A., & Kidwell, S.(2011). Accounting for the effects of biological variability and temporal autocorrelation in assessing the preservation of species abundance. *Paleobiology*, *37*, 332–354.

Toth, A., Lyons, S. K., Barr, W. A., Behrensmeyer, A. K., Blois, J., Bobe, R., et al. (2019). Reorganization of surviving mammal communities after the end-pleistocene megafaunal extinction. *Science*, *365*(6459), 1305–1308.

Uhen, M., Barnosky, A., Bills, B., Blois, J., Carrano, M., Carrasco, M., Erickson, G. M., Eronen, J. T., Fortelius, M., Graham, R. W., Grimm, E. C., O'Leary, M. A., Mast, A., Piel, W. H., Polly, P. D., & Säilä, L. K.(2013). From card catalogs to computers: Databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, *33*(1), 13–28.

Varela, S., Lobo, J., & Hortal, J.(2011). Using species distribution models in paleobiogeography: A matter of data, predictors and concepts. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *310*, 451–463.

Vermeij, G., & Herbert, G.(2004). Measuring relative abundance in fossil and living assemblages. *Paleobiology*, *30*(1), 1–4.

Verstrepen, K.. (2015). *Collaborative filtering with binary, positive-only data*(PhD thesis). Universiteit Antwerpen.

Wang, Y., & Stone, L.(2019). Understanding the connections between species distribution models for presence-background data. *Theoretical Ecology*, *12*(1), 73–88.

Ward, J.(1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Warren, D. L., Matzke, N. J., & Iglesias, T. L.(2020). Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications. *Journal of Biogeography*, *47*(1), 167–180.

Weber, M., Stevens, R., Diniz-Filho, J., & Grelle, C.(2017). Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. *Ecography*, *40*, 817–828.

Yanez-Arenas, C., Guevara, R., Martinez-Meyer, E., Mandujano, S., & Lobo, J.(2014). Predicting species abundances from occurrence data: Effects of sample size and bias. *Ecological Modelling*, *294*, 36–41.

Yu, J., & Dobson, F. S.(2000). Seven forms of rarity in mammals. *Journal of Biogeography*, *27*(1), 131–139.

Zliobaite, I.. (2021). Recommender systems meet species distribution modelling. In *Workshop on perspectives on the evaluation of recommender systems at RecSys'21*. Retrieved from http://ceur-ws.org/Vol-2955/paper7.pdf

Zliobaite, I. (2022). Repository: Recommender systems for fossil species distribution modeling . https://doi.org/10.5281/zenodo.6576483

Zliobaite, I. , Fortelius, M., & Stenseth , N. (2017). Reconciling taxon senescence with the red Queen's hypothesis. *Nature*, *552*, 92 – 95.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Žliobaitė, I. (2022). Recommender systems for fossil community distribution modelling. *Methods in Ecology and Evolution*, *13*, 1690–1706. https://doi.org/10.1111/2041-210X.13916