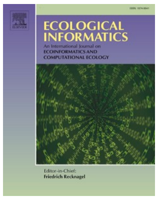
Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Redescription mining for analyzing local limiting conditions: A case study on the biogeography of large mammals in China and southern Asia

Esther Galbrun[a,*], Hui Tang[b], Anu Kaakinen[c], Indrė Žliobaitė[d]

[a] School of Computing, University of Eastern Finland, Technopolis, Microkatu 1, FI-70210 Kuopio, Finland
[b] Department of Geosciences, University of Oslo, P.O. Box 1022, NO-0315 Oslo, Norway
[c] Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, FI-00014, Finland
[d] Department of Computer Science, University of Helsinki, P.O. Box 64, FI-00014, Finland

ARTICLE INFO

Keywords:
Limiting conditions
Climate
Large mammals
Machine learning
Redescription mining
Teeth

ABSTRACT

Identifying and understanding limiting conditions is at the centre of ecology and biogeography. Traditionally, associations between climate and occurrences of organisms are inferred from observational data using regression analysis, correlation analysis or clustering. Those methods extract patterns and relationships that hold throughout a dataset. We present a computational methodology called redescription mining, that emphasizes local patterns and associations that hold strongly on subsets of the dataset, instead. We aim to showcase the potential of this methodology for ecological and biogeographical studies, and encourage researchers to try it.

Redescription mining can be used to identify associations between different descriptive views of the same system. It produces an ensemble of local models, that provide different perspectives over the system. Each model (redescription) consists of two sets of limiting conditions, over two different views, that hold locally. Limiting conditions, as well as the corresponding subregions, are identified automatically using data analysis algorithms.

We explain how this methodology applies to a biogeographic case study focused on China and southern Asia. We consider dental traits of the large herbivorous mammals that occur there and climatic conditions as two aspects of this ecological system, and look for associations between them.

Redescription mining can offer more refined inferences on the potential relation between variables describing different aspects of a system than classical methods. Thus, it permits different questions to be posed of the data, and can usefully complement classical methods in ecology and biogeography to uncover novel biogeographic patterns.

A python package for carrying out redescription mining analysis is publicly available.

1. Introduction

Among the central perspectives in ecology and biogeography is uncovering patterns in the organization of ecological systems and assemblages, and the processes that underlie them (Cox et al., 2020; Dansereau, 1957; MacArthur and Wilson, 1967; Ovaskainen and Abrego, 2020). Contemporary biogeographical studies are data intensive, span increasingly large spatial and temporal scales and require rigorous computational approaches (Pearse and Peres-Neto, 2017). Such analyses typically aim at extracting generic patterns and relations from large observational datasets and highlighting contrasts between different subsets of the data. Most common computational approaches in biogeography (Jongman et al., 1995; Legendre and Legendre, 2012) include correlation analyses, regression analyses, ordination and clustering.

Partitioning techniques, known as clustering, have been part of the toolbox in ecological studies for nearly a century (Kulczynski, 1928). More recently, Kreft and Jetz (2010) and Vavrek (2016) compared clustering methods to identify biogeographic patterns from species distribution data and fossil datasets, respectively. Kreft and Jetz (2010) found that the clusters identified this way were overall similar to the classic primary geographical divisions of the world’s biota, but also exhibit notable differences in the assignment of some subregions, such as, in particular, Madagascar, the Sahara, northern Africa and the Arabian Peninsula.

Ordination techniques aim to reduce the dimensionality of the data while retaining as much information as possible from the original

* Corresponding author.
E-mail address: esther.galbrun@uef.fi (E. Galbrun).

https://doi.org/10.1016/j.ecoinf.2021.101314
Received 7 December 2020; Received in revised form 3 April 2021; Accepted 30 April 2021
Available online 12 May 2021

dimensions. Ordination techniques differ in internal distance measures and complexity of the projection. Examples include general purpose approaches such as Principal Component Analysis (PCA; Pearson, 1901; Hotelling, 1933) or Non-metric Multidimensional Scaling (NMDS; Kruskal, 1964) as well as approaches that are more specifically designed for ecology, such as Outlying Mean Index (OMI; Dolédec et al., 2000) and Ecological Niche Factor Analysis (ENFA; Hirzel et al., 2002).

Regression analysis is broadly used in ecology and biogeography for modelling relationships between variables (see e.g. Ordoñez et al., 2009). Many species distribution models are built on regression (Elith and Leathwick, 2009). New methodological developments aim to take into account spatial (Mellin et al., 2014), multi-scale (Beever et al., 2006) structure of the data or interactions between species (Krapu and Borsuk, 2020).

Combinations of techniques are commonly used as well. For example, Thomas et al. (2019), combine ordination and clustering to investigate how well functional groups explain variance in species traits, while He et al. (2017) identify zoogeographical regions of China through a combination of clustering, ordination and regression analysis. Advanced machine learning techniques are also making their way into biogeographic analysis, Brown et al. (2020), for instance, recently proposed to use support vector machines (SVM) to learn a multi-dimensional boundary between two entities such as populations or species, and examine possible biological overlaps.

Redescription mining, on which we focus here, combines partitioning techniques, such as clustering, and modelling techniques, such as regression. It identifies multiple local models on subsets of data, and automatically generates sets of limiting conditions and the corresponding split of the data. This is where redescription mining departs from most classical analysis methods that identify global models and do not yield explicit and interpretable limiting conditions.

The main idea is to identify two sets of limiting conditions such that, ideally, at any locality they either both hold true or both do not. Thus, redescription mining requires two perspectives of an ecosystem. In this case study, we search for relationships between dental traits of mammals that occur at localities and the climatic conditions of these localities. For example, limiting conditions could require that more than 80% of large herbivores in the region have high crowned teeth and, on the other hand, that the mean annual precipitation and the mean temperature of the warmest quarter in the region be lower than 500 mm and 18 °C, respectively. We would then expect few or no regions satisfying one set of conditions but not the other, that is, having the specified percentage of high-crowned teeth but with rainfall or temperature above the specified thresholds, or vice-versa, satisfying the climatic constraints but having a small percentage of high-crowned teeth.

Here, we tailor redescription mining for analyses in ecology and biogeography. We showcase the potential of this method on a case study looking for associations between the distribution of mammalian dental traits and the climatic conditions of their habitats. Our study focuses on China and southern Asia, which is a pivotal region for biogeographic analyses, due to the complex Asian monsoon climate system and biogeography, affecting the living conditions of approximately one-third of the global human population.

Redescription mining was first introduced as a computational data analysis method in computer science (Ramakrishnan et al., 2004). In addition to algorithmic studies (see references in Galbrun and Miettinen, 2017), this method has been applied, among others, in bio-informatics (Ramakrishnan and Zaki, 2009) and medicine (Mihelčič et al., 2017).

We show how redescription mining can identify biologically meaningful limiting conditions. We also show how those sets of conditions, in the form of redescriptions, can be used to computationally identify or refine zoogeographic units, such as ecoregions.

2. Materials for the case study

China and southern Asia constitute one of the most zoogeographically complex regions in the world due to its diverse environmental gradients, its climatic position, as well as its geological history and spatial inter-connectedness (Ficetola et al., 2017; He et al., 2017). The climate system of China and southern Asia are distinct from any other region in the world.

Variations brought by the Asian monsoon strongly affect the conditions for life in the region (Yamada et al., 2019; Zhao et al., 2010). The plant and animal biomes are diverse and often constitute unique biodiversity hotspots (Z. Tang et al., 2006; Huang et al., 2015). Despite the fact that modern flora and fauna in China and Southern Asia have been strongly fragmented by human activities (He et al., 2018)—which is true for most of the temperate latitudes today—associations between fauna and climate appear to be robust and are subject to active ongoing research (Ficetola et al., 2017; He et al., 2017).

The goal of this case study is to analyze *regional* patterns of association between dental traits of large herbivorous placental mammals and the climatic context of their habitats. Dental proxies used in our analysis are known as dental ecometrics (Eronen et al., 2010; Vermillion et al., 2018; Žliobaitė et al., 2016). Teeth of mammalian herbivores closely reflect the types of plant food their owners can effectively process and convert into energy. Even though each area typically hosts a range of structural types of plant food, different climates will determine which vegetation dominates. Therefore, the distribution of dental traits within faunal communities can provide more robust information about local environmental conditions at the global scale, compared to the presence or absence of specific species, especially of the past ecosystems (Liu et al., 2012).

Previously, we found that global zoogeographic patterns do not directly apply to China and southern Asia (Galbrun et al., 2018). The results suggested complex climate–dental-trait associations prevailing within those spatially compact and climatically unique areas. We hypothesized that the monsoonal climate in these regions may make the conditions attractive for seasonal immigrants from the temperate zones.

2.1. Study region and datasets

The units of our analysis are cells identified by placing a 50 km × 50 km grid over the world map, which we call *localities*. Each locality is characterized by climatic variables as well as variables representing the distribution of dental traits among species occurring at the locality. Functional dental traits are macroscopic, they are defined in such a way that little variation is expected within species, and trait scores can be assigned at the species level (Oksanen et al., 2019). For each locality and each dental trait, we compute the average value over occurring species. We discard localities with fewer than three species, considering that the data in such cells are too sparse for the distribution of dental traits to be informative. In short, our dataset consists of a pair of matrices, *Localities* × *Dental traits* and *Localities* × *Climate* and contains 4416 localities. *Dental traits* and *Climate* comprise respectively 11 and 21 numerical variables. All variables are listed in Table 1.

2.2. Climatic variables

The climate data come from the WorldClim dataset,[1] which builds on extrapolated observations from weather stations. The climatic variables are listed in the right panel of Table 1. We reuse the dataset processed by M. Lawing as reported in (Oksanen et al., 2019). In addition, we considered the net primary productivity (NPP), computed from the mean annual temperature (T˜Y) and total annual precipitation (PTotY) as follows:

[1] http://www.worldclim.org/

$$\mathrm{NPP} = min(3000/(1+exp(1.315-0.119\cdot\mathrm{T}^{\sim}\mathrm{Y})),3000\cdot(1-exp(-0.000664\cdot\mathrm{PTotY}))).$$

We apply a logarithmic transformation to all precipitation variables prior to the analysis with the classical methods. Indeed, these methods rely on identifying linear correlations between variables and are therefore sensitive to the measurement scale. Redescription mining does not require such transformation as it selects thresholds for the limiting conditions independently of the measurement scale.

2.3. Dental trait variables

Species occurrence data come from the list of the International Union for Conservation of Nature.[2] Fig. 1 depicts the number of species occurring at each locality in the study region. We reused the dataset that has been processed by M. Lawing with an extra interpretation of acute lophs as reported in (Oksanen et al., 2019). Dental data have been compiled using the dental trait scoring scheme reported in (Žliobaitė et al., 2016). Teeth are scored by visual inspection, typically of the second upper molar, identifying the presence or absence of specific structural elements and counting specific components, such as cutting edges. The dental variables are listed in the left panel of Table 1. We reuse the scores for species from the study of (Galbrun et al., 2018) with several updates and modifications as follows.

First, we use the average ordinated hypsodonty score instead of binarizing hypsodonty categories, to better align with previous dental ecometric studies. Rather than describing a locality using three variables (fraction of brachydont, mesodont and hypsodont species respectively), we now represent this information with a single variable (averaged hypsodonty). For example, a locality having 30% brachydont, 20% mesodont and 50% hypsodont species, corresponds to mean ordinated hypsodonty of $0.3\cdot 1+0.2\cdot 2+0.5\cdot 3=2.2$. This treatment has been used before in ecometric studies (Eronen et al., 2010; Fortelius et al., 2002; Liu et al., 2012; Žliobaitė et al., 2016). The study of (Galbrun et al., 2018) used binary treatment hoping for higher resolution patterns, but this appeared to be unnecessary.

Table 1
List of the dental traits and bioclimatic variables.

Dental variables	
HYP	Average ordinated hypsodonty
LOP	Average longitudinal loph count
HOD	Average ordinated horizodonty
AL	Fraction of taxa with acute lophs
OL	Fraction of taxa with obtuse lophs
SF	Frac. of taxa with structural fortification of cups
OT	Frac. of taxa with flat occlusal topography
CM	Frac. of taxa with coronal cementum
OO	Frac. of taxa with exclusively obtuse lophs
ETH	Frac. of taxa with thickened enamel
LOPT	Average transverse loph count
Climatic variables	
T$^{\sim}$Y	Mean Annual Temperature
T$^{\sim}$RngD	Mean Diurnal Range
TIso	Isothermality
TSeason	Temperature Seasonality
T^{+}WarmM	Max Temperature of Warmest Month
T^{-}ColdM	Min Temperature of Coldest Month
TRngY	Annual Temperature Range
T$^{\sim}$WetQ	Mean Temperature of Wettest Quarter
T$^{\sim}$DryQ	Mean Temperature of Driest Quarter
T$^{\sim}$WarmQ	Mean Temperature of Warmest Quarter
T$^{\sim}$ColdQ	Mean Temperature of Coldest Quarter
PTotY	Annual Precipitation
PWetM	Precipitation of Wettest Month
PDryM	Precipitation of Driest Month
PSeason	Precipitation Seasonality
PWetQ	Precipitation of Wettest Quarter
PDryQ	Precipitation of Driest Quarter
PWarmQ	Precipitation of Warmest Quarter
PColdQ	Precipitation of Coldest Quarter
NPP	Net Primary Productivity

Temperature and precipitation are measured respectively in degrees Celsius (°C) and in millimeters (*mm*).

Second, we add three dental traits variables, namely *exclusively obtuse lophs* (OO), *thickened enamel* (ETH) and *transverse loph count* (LOPT). The *exclusively obtuse lophs* variable is intended to capture the dental morphology of a generalist, such as a goat. Its value can be derived from the rest of dental traits. For a species, *exclusively obtuse lophs* takes value one if no specialized types of loph-related structures are present (no acute lophs, no structural fortification, no flatness of the occlusal surface). *Thickened enamel* is an experimental trait scored approximately by visual inspection and takes value one if the dental enamel appears to be thicker than regularly seen in molars of other species of a similar size. In this study, the average presence of thickened enamel has a strong taxonomic association with suids. Finally, the *transverse lophs count* is computed in the same way as the *longitudinal lophs count* (LOP) used in our previous study (Galbrun et al., 2018), except that the direction of cutting structures has to span across the tooth row instead of along the tooth row. Both *longitudinal loph count* and *transverse loph count* variables have strong taxonomic associations. The *longitudinal loph count* is high when selenodonts (particularly bovids and cervids) dominate the faunal community. The *transverse loph count* is never dominantly high in faunal communities and increases in the presence of tropical non-Artyodactyls, such as elephants, tapirs or browsing rhinos.

For this analysis, we only use freely available software and libraries. The datasets used in this study along with the scripts for performing the analysis with classical methods as well as with redescription mining, are publicly available at https://github.com/zliobaite/redescription-China.

3. Preamble: classical analysis methods

In order to highlight the perspectives of redescription mining, we first outline patterns and relations that can be produced with the most common classical analysis methods, namely correlation analysis, principal component analysis, regression analysis and clustering. We use implementations provided by the Python SciPy,[3] scikit-learn[4] and Statsmodels[5] libraries.

3.1. Pairwise correlation and scatter plots

Many methods exist for assessing pairwise-relation of numeric variables, the simplest and most popular of which is linear correlation (Pearson correlation coefficient). A correlation coefficient (r) indicates the strength of pairwise association, for example, PWetM and PWetM ($r = 0.995$) vary together, and TSeason and TIso ($r = -0.843$) vary in

[2] https://www.iucn.org/

[3] https://www.scipy.org/, see scipy.cluster.hierarchy

[4] https://scikit-learn.org/stable/, see sklearn.decomposition.PCA, sklearn.cluster.KMeans and sklearn.metrics

[5] https://www.scipy.org/, see statsmodels.api.GLS

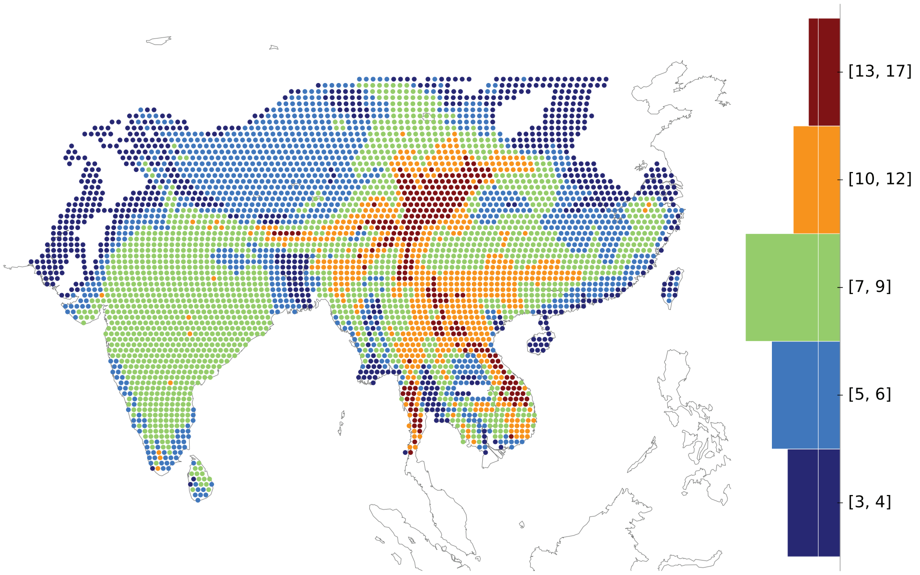


Fig. 1. Map of the species richness. Number of different species of large herbivorous mammals occurring at each locality.

opposite directions. A visual inspection of scatter plots further allows to detect pairs of variables that are strongly related but not in a purely linear way, like PTotY and NPP, or more weakly related in a clearly non-linear way, like T$^{\sim}$DryQ and PWarmQ, for instance. Non-linear methods (such as Spearman rank correlation) or methods for categorical variables are available for quantifying pairwise relationships further, if necessary, but stand-alone correlation analysis does not give a multivariate perspective on data.

3.2. Multivariate projections

Several linear and non-linear methods exist for projecting data into a lower-dimensional space, the most common of which is Principal Component Analysis (PCA) (Hotelling, 1933; Pearson, 1901). The dataset is projected into new dimensions, called the *principal components*, which are positioned orthogonally to each other. For visual analysis, the projection is typically restricted to the first two principal components, i.e. along the two uncorrelated dimensions that preserve the largest amount of variance.

The PCA projection plot in Fig. 2(a) gives an overview of relations between variables. We can identify groups of strongly related variables. For instance, expectedly, monthly and quarterly temperature variables (T^{+}WarmM, T^{-}Cold, T$^{\sim}$WarmQ, etc.) behave in a strongly coordinated manner. We also see that SF and TIso are strongly correlated, and negatively correlated with PSeason.

3.3. Regression models

Regression models are commonly used for making predictions of unobserved variables, as well as summarizing relationships between variables. Various techniques are available for building regression models, starting from single-variable to multi-variable models, from least squares to robust regularized regressions (Hastie et al., 2001), one can also add interaction components, making regression models non-linear.

While PCA belongs to *unsupervised* methods, meaning that no particular perspective or variable is preferred or targeted and the analysis aims at characterizing the structure of the data, regression belongs to *supervised* methods, meaning that particular relationships are assumed and the model detects whether such relationships are present. For instance, the value of PTotY can be estimated accurately from OO (comparatively low Akaike's information criterion (AIC) and high F-statistic values) but models for predicting TIso from the same trait variable do not show a good linear fit. Crucially, the relationships extracted in regression analysis are expected to be valid across most of the dataset, that is, global models are obtained.

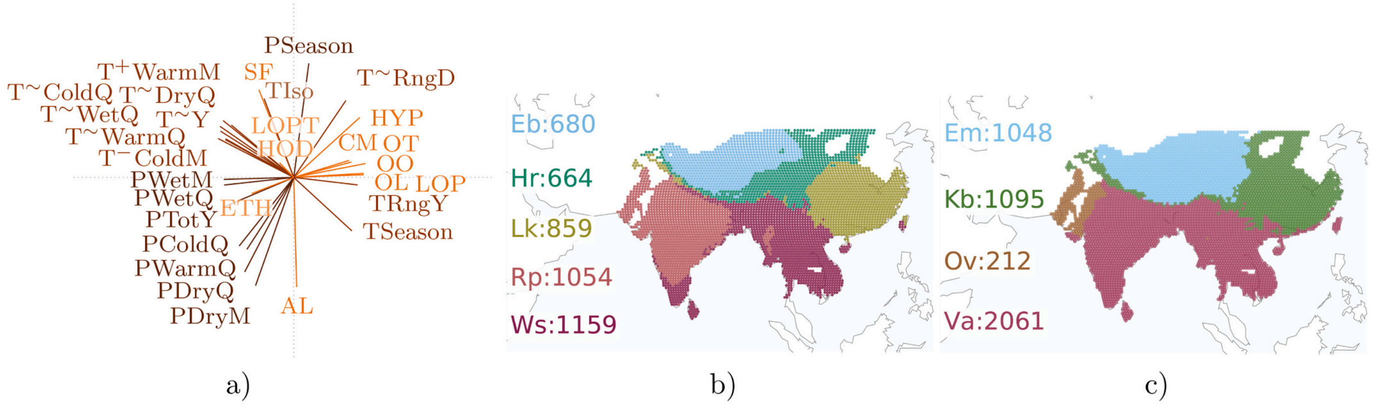


Fig. 2. PCA projection of the variables and maps of clusterings. The variables are projected on the first two components identified by the principal component analysis, considering all variables together (a). The maps show clusterings from *k*-means (b) and HCA with median linkage function (c), both for $k = 5$ clusters. To the left of each map, we list the different clusters, with the number of localities they contain.

3.4. Clustering

Among the many computational techniques available for clustering (Jain and Dubes, 1988), *k*-means and different variants of hierarchical cluster analysis (HCA) are commonly used in ecology.

The *k*-means algorithm (Lloyd, 1982) is an iterative procedure that alternates between assigning data points to the closest cluster center and recomputing the cluster centers. Agglomerative HCA (Ward, 1963) starts with each data point as a distinct cluster. An algorithm then iteratively combines the most similar clusters pairwise, constructing a hierachy of clusters, until a single cluster remains. Practically, the process is often stopped early, when a desired number of clusters is reached. Different criteria for measuring the distance between two clusters lead to variants of the algorithm. Let $\mathrm{d}(x,y)$ denote the distance between two data points x and y, and c_X denote the centroid of cluster X. The distance $\mathrm{D}(U,V)$ between two clusters U and V is defined as follows in different HCA methods, also referred to as *linkage functions*, including:

average a.k.a. Unweighted Pair-Group Method using arithmetic Averages (UPGMA), the average distance between cluster members, $\mathrm{D}(U,V) = \sum_{(u,v)\in U\times V}(u,v)/(|U| \cdot |V|)$

centroid a.k.a. Unweighted Pair-Group Method using Centroids (UPGMC), the distance between cluster centroids, $\mathrm{D}(U,V) = \mathrm{d}(c_U, c_V)$

median a.k.a. Weighted Pair-Group Method using Centroids (WPGMC), a variant of UPGMC weighted by the size of the clusters

For illustration of clustering we use all and variables except HOD and LOPT, which we found to be mostly constant within the focus area. We standardized each variable, i.e. we separately centered and rescaled each variable by subtracting the mean and dividing by the standard deviation. Distances between data points were measured with the ordinary Euclidean distance metric (L^2 norm).

Example clusterings from *k*-means and HCA with median linkage function, both for $k = 5$ clusters are shown in Fig. 2(b) and (c), respectively. Here we do not enforce spatial connectivity of the localities within clusters, however connectivity tends to emerge automatically due to the connectivity of species occurrences and the spatial coherence of climatic trends. Different methods group the localities differently, but some areas emerge across most of the clusterings, regardless of minor variations in the specific localities involved. In particular, localities from the Tibetan plateau and expanding towards the east are often grouped into a cluster (drawn in shades of light blue) and similarly for areas from the Indian subcontinent and the Indochinese peninsula (shades of red and purple), as well as for areas of eastern China (shades of green and brown).

Fig. 3 (top) schematically illustrates how these classical methods operate on a tabular dataset. Clustering identifies different subsets in the data but does not directly offer explanations for why entities are grouped in a particular way and which variables are primarily responsible for the structure. In other words, clustering does not provide models or descriptors of the subsets. Regression or correlation analyses, on the other hand, provide models or descriptors, but they must hold across the whole dataset, without distinctions between subsets. In contrast, redescriptions constitute local models.

4. Redescription mining methodology

The result of redescription mining can be viewed as an ensemble of local models providing multiple perspectives over an ecosystem. The data subsets on which these local models are built can overlap. The local models are not functions, in the sense of a standard regression, but paired collections of limiting conditions, in this case, limits on the climate coupled with limits on the proportion of dental traits among the population of herbivores.

Redescription mining is the process of automatically identifying and statistically evaluating limiting conditions and corresponding data subsets. Different algorithms exist for mining redescriptions. Here we introduce the underlying concepts and one algorithmic approach, which we tailored for biogeographic analyses. See (Galbrun and Miettinen, 2017) for more details about the method. Fig. 3 (bottom) schematically illustrates and summarizes the main concepts of redescription mining.

4.1. Concepts and definitions

With this method, associations are captured as pairs of logical formulas—also known as *queries*—expressing constraints on the values that the variables might take. Each such query defines a subset of localities where the corresponding constraints are satisfied, called the *support* of the query. The algorithmic process constructs pairs of queries, over climate and dental traits variables respectively, such that the two corresponding sets of localities overlap as much as possible. In this way, the method generates alternative descriptions of a subset of localities, in terms of their climatic conditions, on one hand, and of prevailing dental traits, on the other hand, hence the name *redescription*. Queries can be seen as hypotheses about associations between variables, and redescription mining as a process to automatically generate and evaluate those hypotheses.

As a practical example, consider the following query over climatic variables:

$$q_\mathrm{C} = [19.6 \leq T + \mathrm{WarmM} \leq\ ]\ \mathrm{AND}\ [116 \leq PWarmQ].$$

We use the Iverson bracket to specify satisfiability conditions, that is, in our case, the ranges in which the numerical variables must take value. The query above selects localities where the maximum temperature of the warmest month (T^{+}WarmM) is between 19.6 and 38.5 °C and the precipitation of the warmest quarter (PWarmQ) is greater than 116 mm. The support of this query, denoted as $\mathrm{supp}(q_\mathrm{C})$, is the set of localities where the specified temperature and precipitation conditions are satisfied.

Then, a *redescription* is a pair of queries, one over climate variables and one over dental trait variables respectively denoted as q_D and q_C, having similar supports, that is, such that their respective sets of satisfying localities overlap as much as possible. The support of a redescription is the subset of localities at which both queries are satisfied, i.e. the set of localities that meet both the climate as well as the dental conditions. Overloading the notation, we denote the support of a redescription $R = (q_\mathrm{D}, q_\mathrm{C})$ as $\mathrm{supp}(R)$, which is such that

$$supp(R) = supp(q_\mathrm{D}) \cap supp(q_\mathrm{C}).$$

The *accuracy* of a redescription is a measure of the validity of the relationship across the dataset. The accuracy could be assessed using any similarity measure between sets. The Jaccard coefficient (Jaccard, 1901) is generally used for this purpose because it is intuitive and symmetric, in the sense that the two compared sets are exchangeable. Formally, the Jaccard coefficient is defined as

$$J(R) = \frac{|supp(q_\mathrm{D}) \cap supp(q_\mathrm{C})|}{|supp(q_\mathrm{D}) \cup supp(q_\mathrm{C})|}.$$

Informally, we are trying to maximize the number of localities where both queries are satisfied while minimizing the number of localities where only one of them is. To assess the statistical significance, we compute a that indicates how likely it is that the support of the redescription is as large or larger than observed, given the size of the support of the two queries it consists of, assuming the queries are independent.

4.2. Analysis procedure and parameter settings

Multiple algorithms have been proposed for finding accurate and statistically significant redescriptions. In this study, we use the algorithm (Galbrun and Miettinen, 2012), which is a greedy algorithm in the sense that it makes a locally optimal choice at each iteration. In the initialization phase, the algorithm tests all variable pairs, in our case each dental variable with each climatic variable, aiming to form simple

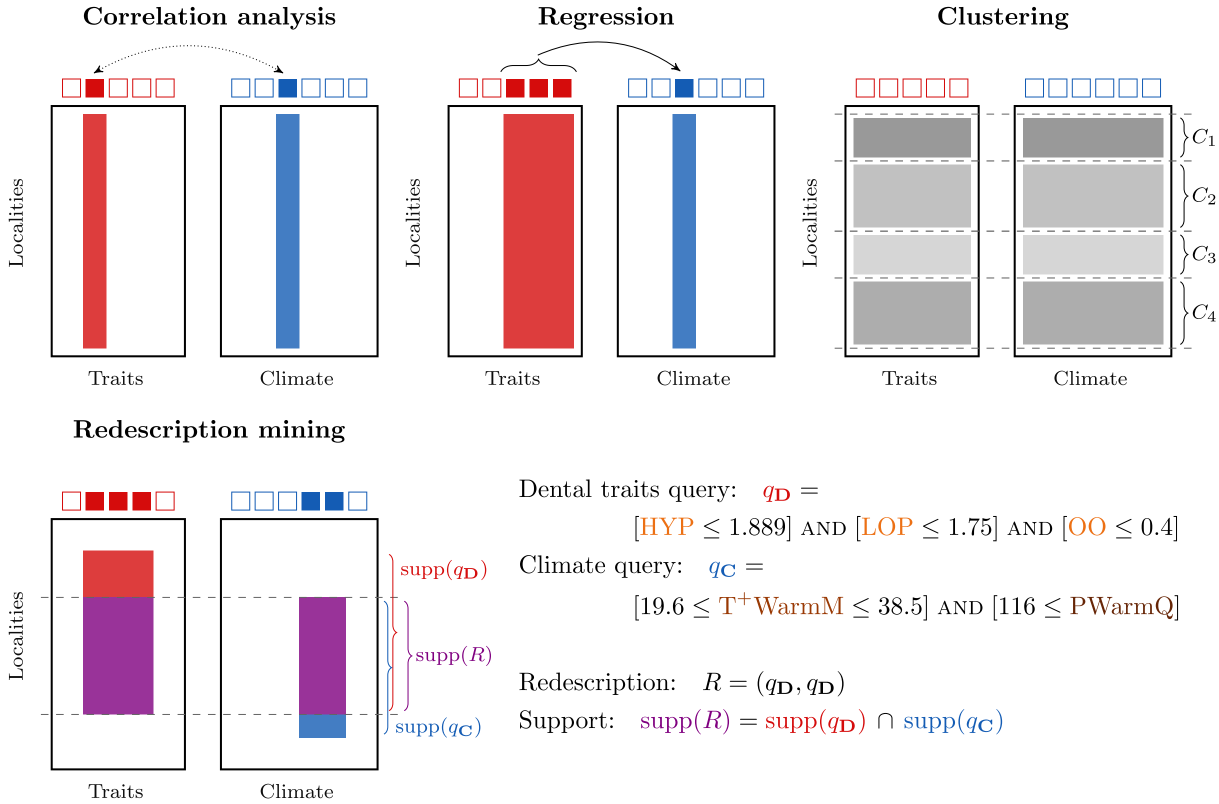

Correlation analysis
Regression
Clustering
Localities
Traits
Climate
C_1
C_2
C_3
C_4
Redescription mining
$\mathrm{supp}(q_{\mathbf{D}})$
$\mathrm{supp}(R)$
$\mathrm{supp}(q_{\mathbf{C}})$
Dental traits query: $q_{\mathbf{D}} =$
$[\mathrm{HYP} \leq 1.889]$ AND $[\mathrm{LOP} \leq 1.75]$ AND $[\mathrm{OO} \leq 0.4]$
Climate query: $q_{\mathbf{C}} =$
$[19.6 \leq \mathrm{T^{+}WarmM} \leq 38.5]$ AND $[116 \leq \mathrm{PWarmQ}]$
Redescription: $R = (q_{\mathbf{D}}, q_{\mathbf{D}})$
Support: $\mathrm{supp}(R) = \mathrm{supp}(q_{\mathbf{D}}) \cap \mathrm{supp}(q_{\mathbf{C}})$

Fig. 3. Schema of classical methods we consider (top) and redescription mining, including a summary of important notations (bottom). Classical methods can be separated into variable-centered approaches (including correlation analysis and regression analysis) and locality-centered approaches (including clustering). Redescription mining aims to combine these two types of approaches.

redescriptions. In the extension phase, the algorithm then iterates over these basic redescriptions and extends them, aiming to improve the accuracy of the redescription. Specifically, generates redescriptions by appending new variables to the current queries, at each step keeping the best candidates for further extension.

We performed the analysis using Siren,[6] an interface that allows to automatically generate redescriptions with various algorithms, including, and to visualize, cluster and interactively edit the redescriptions (Galbrun and Miettinen, 2018).

The method requires manually setting several parameters, described in more details in the user guide.[7] In particular, about half a dozen parameters allow to set thresholds on the size of the support of the output redescriptions and to control the length and complexity of their queries.

We required that at least 1% of localities satisfy both queries (MinSuppIn) and that at least 30% of localities satisfy neither of the queries (MinSuppOut). In other words, the intersection of the supports of the two queries (the support of the redescription) and their union were required to contain at least 1% and at most 70% of all localities. This is an inclusive choice, not overly restrictive, that aims at capturing local patterns. Increasing the upper threshold further would jeopardize the local aspect of the analysis, and would lead to something more akin to non-linear regression. For a redescription to be informative, its support should neither be too large nor too small, and we found these thresholds to provide a good balance, and small variations in these parameters did not impact the results much.

We used two different setups when running the ReReMi algorithm. In the first run, we allowed only conjunctive queries on both sides (i.e. we explicitly forbid the use of 'OR') and restricted the number of variables to three dental variables and two climate variables. In the second run, we allowed dental queries to involve disjunctions, and climate queries to contain up to three variables, but tightened the requirement of accuracy gain. Specifically, under this constraint, a candidate query can be extended by automatically adding the next variable only if the accuracy, as measured by the Jaccard coefficient, increases by a least 0.1. The goal is to obtain interpretable, not overly complex (long) queries. This can be achieved either explicitly, by limiting the number of variables and the operators used in the queries, as in the first run, or implicitly, by allowing increased complexity only if it brings substantial improvement in terms of accuracy, as in the second run.

4.3. Selecting individual redescriptions for further analyses

Redescription mining typically outputs a large number of redescriptions, each holding on a subregion within the dataset. Subregions can overlap, and the same subregion can potentially be described by different variables. Analysts might manually sift through individual redescriptions. However, it is not practical to analyze large collections of redescriptions, since many of them contain similar information. Therefore, computational means are needed to remove redundant (very similar) redescriptions and identify the most informative (distinct) patterns.

In this study we approach this challenge in three ways. First, we rank and filter redescriptions automatically using accuracy and redundancy measures. Among the top-listed redescriptions, we pick a few pairs for further analysis by visually inspecting maps of the corresponding subregions. We also analyze the top-listed redescriptions as a group, by means of clustering, allowing us to identify coherent computational ecoregions for the study area. In other words, we perform our analysis and reach conclusions through a combination of automated and manual

[6] http://cs.uef.fi/siren/main/

[7] http://cs.uef.fi/siren/help/

processing.

The first run, with strict explicit constraints, generated 271 redescriptions while the second run, with stringent threshold on accuracy gains, generated 188 redescriptions. Either run took about 50 min to complete on a commodity laptop.

We filter the two collections separately, ranking the redescriptions by decreasing accuracy and removing any redescription having more than 90% of its support in common with a higher-ranked one. That is, a redescription R_x is removed from the set of results if it contains a more accurate redescription R_y such that

$$\frac{|supp(R_x) \cap supp(R_y)|}{min(|supp(R_x)|,|supp(R_y)|)} > 0.9.$$

We then inspect more closely the top-ten remaining redescriptions from both lists. We denote the ten redescriptions produced by the first run, i.e. using only conjunctions, and ordered by decreasing accuracy as R1.1–R1.10. Similarly, the ten redescriptions produced by the second run, i.e. under the stricter improvement requirement, and ordered by decreasing accuracy are denoted as R2.1–R2.10. All the selected redescriptions have s close to zero, without correction for multiple testing which is not yet possible with existing methods.

In summary, the twenty selected redescriptions were obtained using automated processes driven primarily by accuracy with the second run yielding more compact but somewhat less accurate redescriptions than the first.

4.4. Using redescriptions as building blocks for identifying new ecoregions

While individual redescriptions and the associated limiting conditions can be analyzed in isolation (Section 5.2), they can also be used in combination as lenses to characterize ecosystems (Section 5.1). Conceptually, each redescription can be thought of as a basic ingredient. Each locality then can be described by a recipe, that involves some of these ingredients (redescriptions that hold true at that locality) but not others. We can then find similar localities in terms of their redescription profiles and denote them as distinct ecoregions. The procedure is as follows.

Each locality is represented by a binary vector recording whether or not the corresponding redescription holds at the locality, which we refer to as the *support membership vector*. The distance $\mathrm{h}(u,v)$ between two localities is measured as the Hamming distance, i.e. the number of mismatches, between their respective support membership vectors. In other words, the distance between localities u and v is the number of redescriptions that hold at either of the two localities but not both. The distance is zero if the localities satisfy exactly the same redescriptions.

Clusters are then formed by applying a hierarchical agglomerative procedure to the support membership vectors. As with standard hierarchical clustering methods, we obtain different variants depending on how the distance between clusters is measured, and hence how the next pair of clusters to merge is selected. The distance $\mathrm{D}(U,V)$ between two clusters U and V is defined as follows in the different redescription clustering methods:

sizes the maximum distance between cluster members, i.e. $\mathrm{D}(U,V) = \max_{(u,v)\in U\times V}\mathrm{h}(u,v)$. Ties are broken in favor of pairs of clusters having similar sizes.

ones the maximum distance between cluster members, i.e. $\mathrm{D}(U,V) = \max_{(u,v)\in U\times V}\mathrm{h}(u,v)$. Ties are broken in favor of pairs of clusters sharing more positive matches, first, and having similar sizes, second.

wdist the sum of distances between cluster members, i.e. $\mathrm{D}(U,V) = \sum_{(u,v)\in U\times V}\mathrm{h}(u,v)$, directly taking into account the sizes of the clusters.

Because the clusters are generated based on which redescriptions the localities support, the redescriptions that are most represented within each cluster provide a characterization for it. In other words, the queries of the redescriptions can be used to understand what are the properties that lead to localities being grouped together into a cluster. Each cluster can be interpreted as a computationally identified ecoregion.

5. Case study: biogeographic analysis with redescription mining

The goal of this case study is to illustrate the type of insights and interpretations that can potentially be obtained from redescriptions. We first explain how an analysis can be performed at the ecosystem level, using redescriptions as ingredients. Then, we focus on a few selected redescriptions to show what type of information they can capture.

5.1. Computationally identifying ecoregions with redescription summaries

We obtain summaries in terms of the twenty most accurate redescriptions with the different clustering variants and for $k = 3$, 5 and 7 clusters. These clusterings reflect limiting conditions in terms of dental traits and climate variables. We focus on the summary obtained with the widst clustering variant, as it accounts for cluster sizes in a natural way, and is therefore fairly interpretable. The summary obtained by setting the number of clusters to 5, which gives the best compromise between number of clusters and total distance, is shown in Fig. 4.

The left panel of Fig. 4 shows a map of the resulting geographic clusters. As above with the support of redescriptions, the clusters tend to span over contiguous localities, not because we enforce spatial connectivity but, rather, as a consequence of autocorrelation within the variables. Since the clusters summarize the interplay between the support of multiple redescriptions, individual redescriptions are not expected to match the boundaries of any single cluster.

The results suggest generally similar distinct biogeographical regions as the clustering analysis based on raw dental traits and climate data (Section 3), such as the Tibetan Plateau, East China and India. However, they exhibit a lower similarity between India and Southeast Asia and between the Tibetan Plateau and northern China, but a greater similarity between southern China and Southeast Asia and between southern China and northern China. A much finer spatial structure over southern China and Southeast Asia is captured, which seems to correspond well to the distribution of plant relicts found in these regions (Huang et al., 2015).

The right panel of Fig. 4 shows how the localities supporting the redescriptions are distributed among these cluster regions. This can be used to look up the redescriptions that are most represented within a cluster (darker cells) and understand the reasons that led to the cluster being formed. For instance, redescriptions R2.1 and R1.2 are very specific to cluster **A**, which corresponds to the Tibetan plateau. The Tibetan Plateau and surrounding regions, indeed, have been highlighted as one of the most complex and distinct biotas on Earth (He et al., 2020), that also underwent striking changes over time. Redescriptions R1.8, R2.6 and R1.10 and R1.5 are particularly well represented in cluster **B**, while redescriptions R2.5 and R1.4 are represented in both **B** and **C**, as well as cluster **D** to a lesser extent. Most of the remaining redescriptions are represented in cluster **C**, as well as cluster **D**, cluster **E**, or both.

The geographic clusters of the redescriptions summary can be thought of as computational ecoregions. For comparison, the terrestrial ecoregions[8] as defined by Olson and Dinerstein (2002) are plotted in Fig. 5 and listed in Table 2. While the original mapping is primarily based on vegetation, zoogeographically adjusted variants (Holt et al., 2013) offer by and large the same conclusions with somewhat more pronounced separation between India and southern Asia.

We observe some correspondences between the terrestrial ecoregions and the clustering that emerges from the support of the redescriptions. These results suggest that the patterns extracted automatically from the dental traits distribution and climatic variables, without any geospatial information, closely correspond to manually defined terrestrial ecoregions. Redescription cluster **A** closely matches

[8] http://maps.tnc.org/gis_data.html

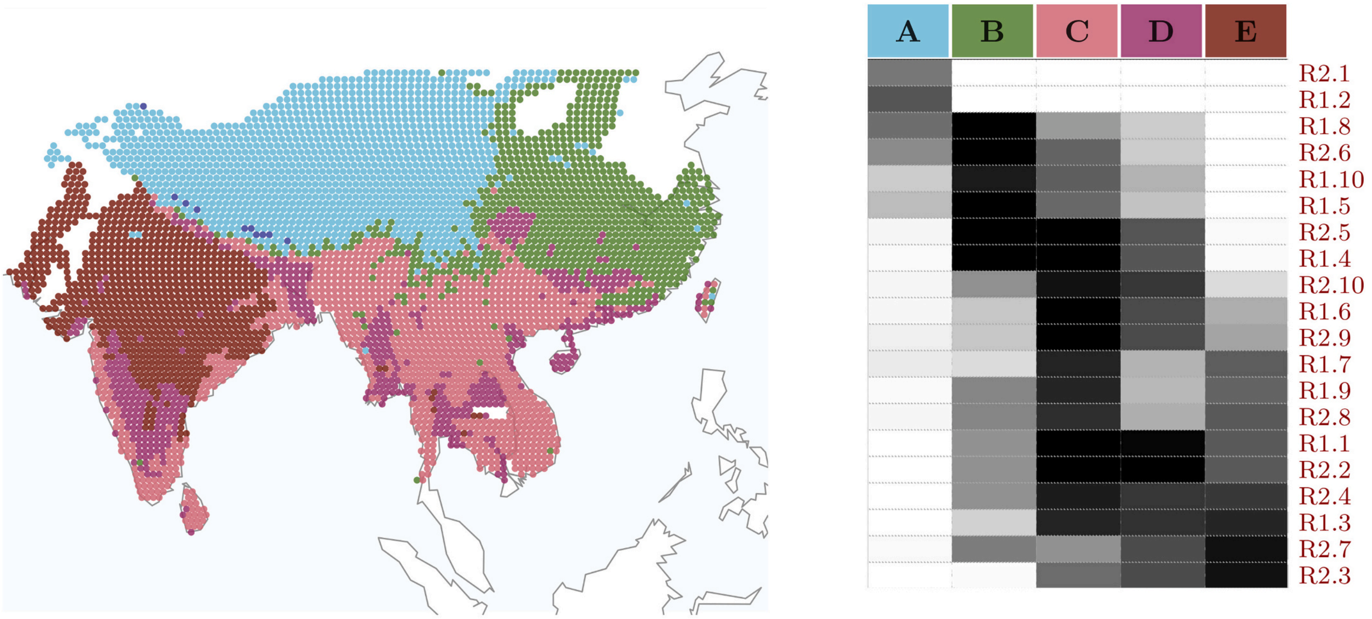


Fig. 4. Redescription summary. Redescription-based clustering with the wdist variant and $k = 5$ clusters. The left-hand side panel shows a map of the five cluster regions formed by the supports of the top ten redescriptions from both runs. The table in the right-hand side panel shows the repartition of the supports of the redescriptions across these cluster regions. Each column of the table corresponds to one of the cluster regions and each row corresponds to a redescription. The shade of the cells indicates the fraction of localities from the region belonging to the support of the redescription, with black cells meaning that the entire region belongs to the support of the redescription.

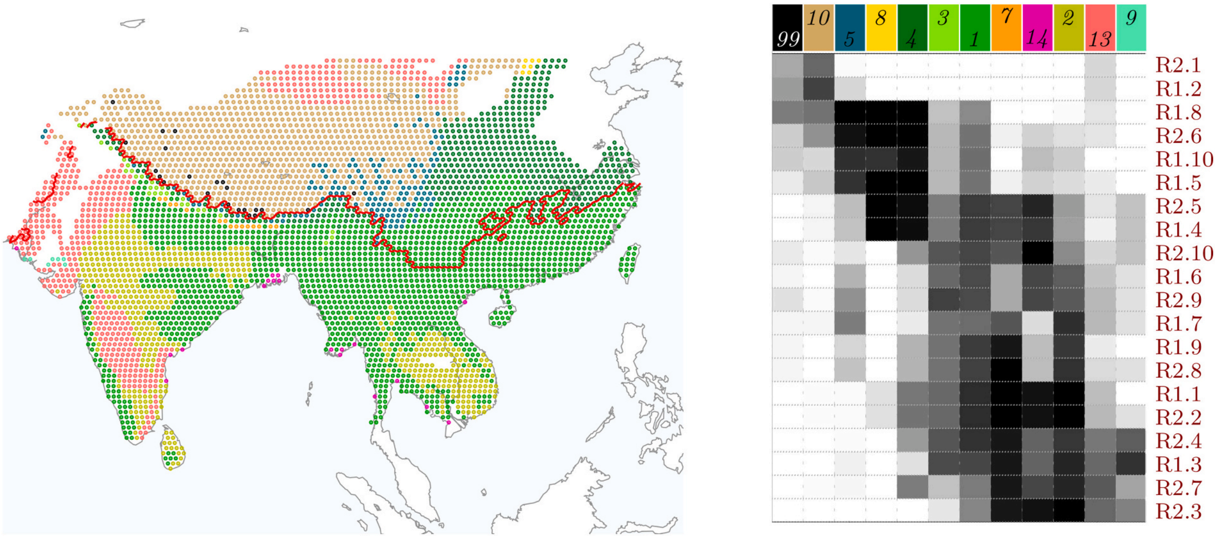


Fig. 5. Distribution of the twenty redescriptions across terrestrial ecoregions. The left-hand side panel shows a map of the terrestrial ecoregions in the study region, with the border between the Palearctic (north) and Indomalaya (south) biogeographic realms as a red line (Olson and Dinerstein, 2002). The table in the right-hand side panel shows the repartition of the support of the redescriptions across these ecoregions. Similarly as in Fig. 4, each column of the table corresponds to one of the ecoregions and each row corresponds to a redescription. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the *Montane Grasslands and Shrubland* ecoregion (ecoregion 10 in Fig. 5 and Table 2). Redescription cluster **B** closely captures *Temperate Broadleaf and Mixed Forests* (ecoregion 4) and *Temperate Conifer Forests* (ecoregion 5). The narrow band of temperate forest along the southern slope of the Himalayas is especially well captured. Redescription cluster **C** largely matches *Tropical and Subtropical Moist Broadleaf Forests* (ecoregion 1) except some misses in India. Redescription cluster **E** mainly covers inland India and matches *Tropical and Subtropical Moist* as well as *Dry Broadleaf Forests* (ecoregions 2 and 1), while redescription cluster **D** collects many isolated patches nearby coasts, to the exclusion of coasts corresponding to *Deserts and Xeric Shrubland* (ecoregion 13), to *Mangroves* (ecoregion 14), and to *Broadleaf Forest ecoregions* (ecoregions 1, 2 and 4).

Comparison with the map of species richness of large herbivorous mammals (Fig. 1) reveals a good overlap of cluster **D** with the regions showing low number of species over Southeast Asia, Bangladesh and the southern coast of China. This implies that cluster **D** may emerge due to a lack of data. However, Cluster **D** also appears to be visually similar to the distribution of *plant relicts*, i.e. “plant groups that were once widespread in the Northern Hemisphere but are now restricted to some small

Table 2
List of the terrestrial ecoregions (Olson and Dinerstein, 2002).

1	Tropical and Subtropical Moist Broadleaf Forests
2	Tropical and Subtropical Dry Broadleaf Forests
3	Tropical and Subtropical Coniferous Forests
4	Temperate Broadleaf and Mixed Forests
5	Temperate Conifer Forests
6	Boreal Forests/Taiga
7	Tropical and Subtrop. Grasslands, Savannas and Shrublands
8	Temperate Grasslands, Savannas and Shrublands
9	Flooded Grasslands and Savannas
10	Montane Grasslands and Shrublands
11	Tundra
12	Mediterranean Forests, Woodlands and Scrub
13	Deserts and Xeric Shrublands
14	Mangroves
98	Inland Water
99	Rock and Ice

isolated areas", in southern China (Huang et al., 2015). Therefore, this represents a region (corresponding to cluster **D** in our results) with a unique climate-vegetation association, which cannot be observed in other places anymore nowadays. Cluster **D** also relates to the potential distribution of Savannahs in Asia (Fig. 1 in Ratnam et al., 2016) and its spatially fragmented nature is similar to the distribution of high mammalian diversity regions in Asia (Brum et al., 2017). These lines of evidence may explain the lack of spatial connectivity of cluster **D**, and indicate that the seemingly randomly distributed regions of cluster **D** are more likely to arise from their unique biogeographical features as reflected by their climate and dental traits together. We emphasize that cluster **D** does not emerge by doing clustering analysis on dental traits or climate variables alone, highlighting the potential of redescription mining for recognizing unique biogeographical regions (e.g. biodiversity hotpots or Savannahs).

This part of our analysis has a close connection to the recent work of He et al. (2017). The main distinction, apart from the fact that we include southern Asia in addition to China, is in the source information. He et al. (2017) used species occurrence lists, while we primarily relied on functional dental traits. This way, our species coverage is narrower, but hopefully provides a direct biomechanical link, with vegetation as an interface between plants and the animals that eat them. Functional dental traits primarily relate to limiting rather than average climatic conditions for herbivores (Žliobaitė et al., 2016). Given those methodological differences it is reassuring to observe a general match in the prominence of the Tibetan plateau and the East–West division.

5.2. Insights from individual redescriptions

We now take a closer look at a selection of individual redescriptions. We analyze individual redescriptions from two runs. We selected the most accurate matching redescriptions from each run. As visible from the support maps in Fig. 6, the selected redescriptions characterize distinctive regions.

Redescriptions R2.1 and R1.2 cover the Tibetan plateau. Redescription R2.1 requires longitudinal loph count (LOP) to be close to maximum and no thickened enamel (ETH), which in the context of our experimental scoring relates to the absence of suids. Redescription R1.2 requires obtuse lophs (OL) to be close to maximum, indicating generalist herbivory (Oksanen et al., 2019), and structural fortification (SF) to be very low, which hints towards seasonal environments lacking humid woodlands (Žliobaitė et al., 2018), as well as low proportion of thickened enamel (ETH) as before. From the climatic perspective, redescription R2.1 prescribes low temperatures in the warmest quarter ($T^{\sim}$WarmQ) and low annual precipitation (PTotY), while redescription R1.2 prescribes a low mean annual temperature ($T^{\sim}$Y) and high but not extreme seasonality of the temperature (TSeason). Indeed, these redescriptions align with harsh seasonal environments in combination with generalist dental morphologies. Note that R2.1 seems to capture even harsher and continental climates than R1.2. The support of R1.2 is smaller than the support of R2.1 and does not cover part of the south-eastern Tibetan plateau.

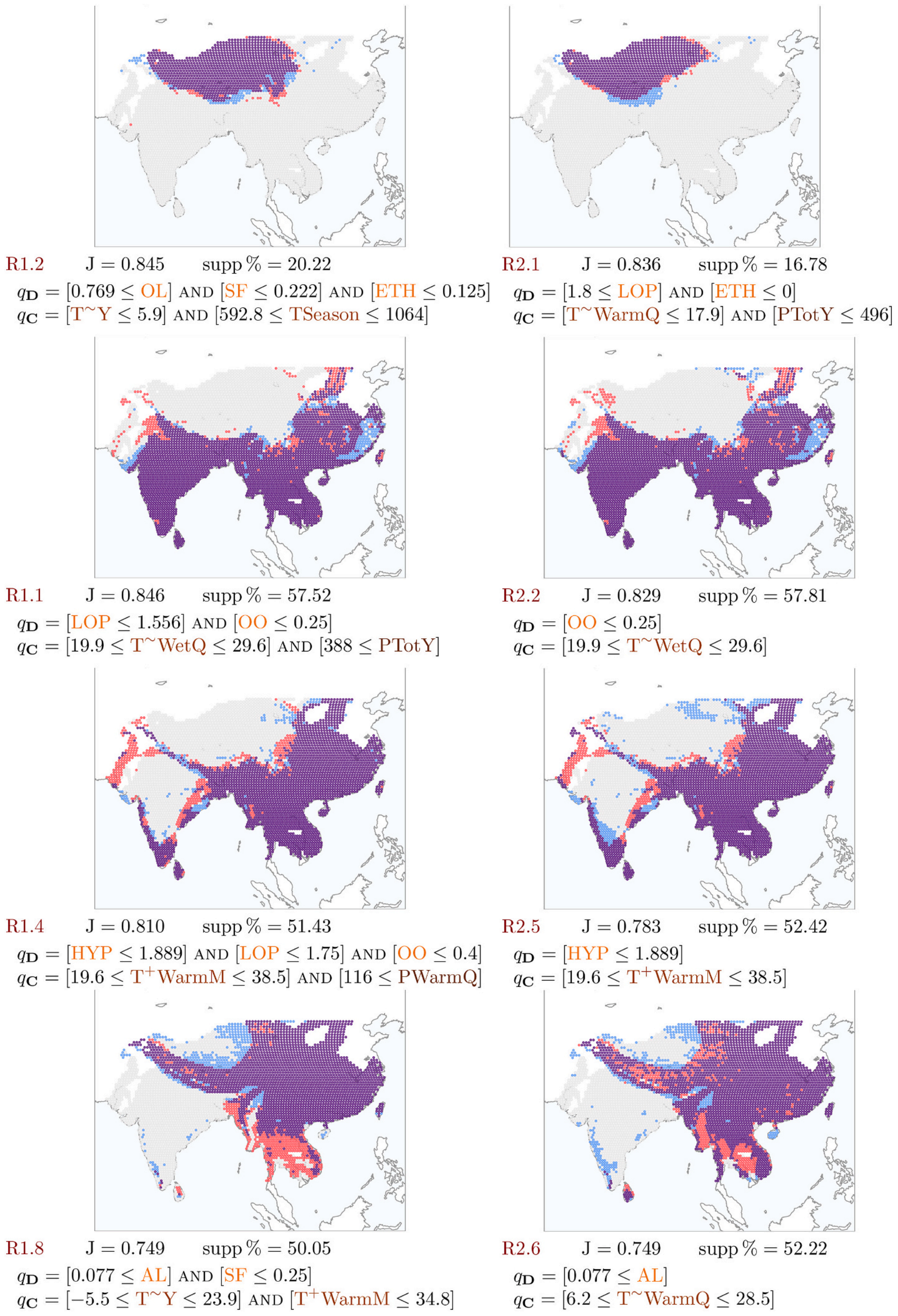
Redescriptions R2.2 and R1.1 cover the majority of the continental part of the Indomalaya biogeographic realm, yet have tightly restricted queries both from the dental and climatic perspectives, which associate with relatively wet and woody habitats. Both dental queries require a low share of exclusively obtuse lophs (OO), which suggests a dominance of browsers in closed habitats and is in line with dominant closed woody vegetation in that region. The dental query of the second redescription excludes extreme high longitudinal loph count (LOP), which suggests a combination of selenodont and non-selenodont teeth, which is expected in the context of near-tropical woody vegetation. From the climatic perspective, both redescriptions require the temperature of the wettest quarter ($T^{\sim}$WetQ) to be warm, but not too hot, suggesting the presence of an extremely favorable growing season. The second redescription, R1.1, further requires annual precipitation (PToY) not to be too low. Overall, these redescriptions and their support regions hint towards an accommodating environment, which does not require extremely specialized teeth and supports a high richness of herbivore species (cf. Fig. 1).

Redescriptions R2.5 and R1.4 describe a subset of the Indomalaya biogeographic realm, excluding inland India but extending north into China. Both dental queries require hypsodonty (HYP) to be relatively low. Redescription R1.4, further requires a low loph count (LOP) and a low share of obtuse lophs (OO), similarly to the previous pair of redescriptions covering the Indomalaya realm (R2.2 and R1.1). Both climate queries require temperatures of the warmest month (T^{+}WarmM) to range from rather mild to quite hot (from ca. 20 °C to ca. 40 °C). Redescription R1.4 further constrains precipitation of the warmest quarter (PWarmQ), excluding extreme dryness. Curiously, the two redescriptions cover costal areas of India and the foothills of the Himalayas, but not central India, where hypsodonty tends to be higher.

Redescriptions in the last pair (R2.6 and R1.8) show a curious spatial pattern. They cover primarily mainland East China and southern Asia, extending into a narrow strip spanning across the slope of the Himalaya mountains, without ever including the top (Tibetan plateau) nor the bottom (central India) of the mountain range (cf. bottom row of Fig. 6). The dental queries of both redescriptions include acute lophs (AL). The specified range of values is broad, allowing all except total and near-absence, and is thus not particularly informative. However, high proportions of acute lophs generally indicate seasonal temperate environments with abundant woody cover (Oksanen et al., 2019), of deciduous forests in particular. The second redescription includes a constraint to low proportion of structurally fortified molars (SF). Structural fortification is generally a characteristic of tropical woody environments, and often comes along with high hypsodonty (Žliobaitė et al., 2018). Only temperature variables appear in the climate queries of both redescriptions. The first redescription allows a wide range of temperatures during the warmest quarter ($T^{\sim}$WarmQ), down to a rather cold lower bound (6 °C). The second redescription instead involves the mean annual temperature ($T^{\sim}$Y), also allowing a wide range of values, down to rather low values (−5.5 °C).

The last two pairs of redescriptions (R2.5 and R1.4) and (R2.6 and R1.8) are quite similar in terms of their geographic coverage, with the latter pair almost eschewing the Indomalayan realm while having a much broader coverage along the Himalayan slope and more coverage in more northern parts of the region. On the climate side, both pairs emphasize the warmest periods of the year, with the latter pair having a lower threshold for the warmest temperature. In terms of traits the first pair emphasizes (lack of extreme) durability via hypsodonty (HYP), while the second pair emphasizes the cutting capacity via acute lophs (AL).

Overall, an in-depth analysis of every obtained redescription would normally be infeasible. Indeed, our runs produced a total of 459



R1.2 J = 0.845 supp % = 20.22

$q_{\mathrm{D}} = [0.769 \leq \mathrm{OL}]$ AND $[\mathrm{SF} \leq 0.222]$ AND $[\mathrm{ETH} \leq 0.125]$

$q_{\mathrm{C}} = [\mathrm{T}^{\sim}\mathrm{Y} \leq 5.9]$ AND $[592.8 \leq \mathrm{TSeason} \leq 1064]$

R2.1 J = 0.836 supp % = 16.78

$q_{\mathrm{D}} = [1.8 \leq \mathrm{LOP}]$ AND $[\mathrm{ETH} \leq 0]$

$q_{\mathrm{C}} = [\mathrm{T}^{\sim}\mathrm{WarmQ} \leq 17.9]$ AND $[\mathrm{PTotY} \leq 496]$

R1.1 J = 0.846 supp % = 57.52

$q_{\mathrm{D}} = [\mathrm{LOP} \leq 1.556]$ AND $[\mathrm{OO} \leq 0.25]$

$q_{\mathrm{C}} = [19.9 \leq \mathrm{T}^{\sim}\mathrm{WetQ} \leq 29.6]$ AND $[388 \leq \mathrm{PTotY}]$

R2.2 J = 0.829 supp % = 57.81

$q_{\mathrm{D}} = [\mathrm{OO} \leq 0.25]$

$q_{\mathrm{C}} = [19.9 \leq \mathrm{T}^{\sim}\mathrm{WetQ} \leq 29.6]$

R1.4 J = 0.810 supp % = 51.43

$q_{\mathrm{D}} = [\mathrm{HYP} \leq 1.889]$ AND $[\mathrm{LOP} \leq 1.75]$ AND $[\mathrm{OO} \leq 0.4]$

$q_{\mathrm{C}} = [19.6 \leq \mathrm{T}^{+}\mathrm{WarmM} \leq 38.5]$ AND $[116 \leq \mathrm{PWarmQ}]$

R2.5 J = 0.783 supp % = 52.42

$q_{\mathrm{D}} = [\mathrm{HYP} \leq 1.889]$

$q_{\mathrm{C}} = [19.6 \leq \mathrm{T}^{+}\mathrm{WarmM} \leq 38.5]$

R1.8 J = 0.749 supp % = 50.05

$q_{\mathrm{D}} = [0.077 \leq \mathrm{AL}]$ AND $[\mathrm{SF} \leq 0.25]$

$q_{\mathrm{C}} = [-5.5 \leq \mathrm{T}^{\sim}\mathrm{Y} \leq 23.9]$ AND $[\mathrm{T}^{+}\mathrm{WarmM} \leq 34.8]$

R2.6 J = 0.749 supp % = 52.22

$q_{\mathrm{D}} = [0.077 \leq \mathrm{AL}]$

$q_{\mathrm{C}} = [6.2 \leq \mathrm{T}^{\sim}\mathrm{WarmQ} \leq 28.5]$

Fig. 6. Focus maps of example redescriptions. Localities that support both queries, only the dental trait query and only the climate query, are drawn in purple, in red and in blue, respectively. For each redescription, we list the query over dental traits variables (q_{D}), the query over bioclimatic variables (q_{C}), the accuracy of the redescription (J) as well as the size of its support as a percentage of the total number of localities (supp %). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

redescriptions. Each redescription represents one local perspective towards an ecosystem. One can select individual redescriptions for analysis using quantitative criteria, or use them together as elements in structural analyses of ecosystems. This type of analysis also has potential in studies of past and future ecosystems, where some elements can be expected to vary over time. Decomposing an ecosystem into such functional elements might help, for instance, investigate which aspects of the system are changing over time and which aspects remain constant.

6. Conclusions

Redescription mining is a methodology for extracting local patterns between two perspectives over the same system. It can be seen as a hybrid of regression modelling and cluster analysis. Indeed, it delineates subsets of the data, similarly to clustering, and also captures relationships between variables, similarly to regression. Some descriptions might be generic and hold across a large number of localities, whereas other descriptions might be very specific and hold only at few localities.

In our case study, we analyze dental traits and climate variables in China and southern Asia via redescription mining. We show that individual redescriptions allow to identify spatial associations between dental traits and climate variables, while redescription summaries (i.e. clusterings based on the redescriptions) can delineate distinct biogeographic areas within this region. We show how an ecosystem level analysis can be carried out using redescriptions as elements, and then zoom into selected redescriptions to show how they can capture ecological limiting conditions.

The results based on redescriptions reveal a finer spatial structure over southern China and Southeast Asia, which seems to correspond well to plant relicts found in these regions. In contrast, the results of the classical clustering focus on the differences within the Tibetan Plateau. These discrepancies highlight the potential added value of using redescriptions-based clusters to delineate biologically meaningful ecoregions with finer structure.

Different from regression methods that require strong assumptions on the shape of the association across the whole data set (e.g. linear or logarithmic), redescription mining allows a broader exploration of different associations for different subsets of the data, that can not be detected by classical methods. Redescription mining searches for pairs of descriptions that intersect in their areas of validity. In the ecological sense, a redescription automatically extracts and pairs collections of limiting conditions, such that if one collection of conditions is satisfied, the other is also very likely to be satisfied. Since redescription mining works by automatically identifying limiting conditions from two perspectives, it naturally lends itself to ecological analyses, where limiting conditions often play a central role.

Through redescriptions, localities can be characterized in different ways in terms of the available variables, e.g. specific occurring species, species richness, vegetation types, average climate or elevation. The methodology is not limited to finding associations across space, when the considered objects have geospatial coordinates like the localities considered in this study. It can also be used to identify associations across time or biological organisms. It can be applied to different types of variables describing various aspects of an ecosystem, such as species abundance, plant traits, human disturbances, etc., and to other regions, depending on the research questions. We believe that redescription mining offers an interesting complementary tool for biogeographic and ecological analyses.

One can select individual redescriptions for analysis using quantitative criteria, or use them all together as elements for structural analyses of ecosystems. This type of approach also has potential for studying past and future ecosystems, for instance to help tell apart aspects of the system that are changing over time from those that remain constant.

Author contributions

HT, IŽ and EG prepared the data. EG performed the analysis with the different methods. All authors analyzed the results. EG and IŽ wrote the first version of the paper. All authors contributed to editing the manuscript and gave final approval for publication.

Source code and data availability

For this analysis, we only use freely available software and libraries. The datasets used in this study along with the scripts for performing the analysis with classical methods as well as with redescription mining, are publicly available at https://github.com/zliobaite/redescription-China.

Declaration of competing interest

None.

Acknowledgements

Research leading to these results was partially funded by the Academy of Finland (grant no. 314803 to I.Ž. and grant no. 316799 to A.K.) and the Research Council of Norway (EMERALD project, grant no. 294948, H.T.).

References

Beever, E., Swihart, R., Bestelmeyer, B., 2006. Linking the concept of scale to studies of biological diversity: evolving approaches and tools. Divers. Distrib. 12 (3), 229–235. https://doi.org/10.1111/j.1366-9516.2006.00260.x.

Brown, M., Holland, B., Jordan, G., 2020. hyperoverlap: detecting biological overlap in n-dimensional space. Methods Ecol. Evol. 11 (4), 513–523. https://doi.org/10.1111/2041-210X.13363.

Brum, F.T., Graham, C.H., Costa, G.C., Hedges, S.B., Penone, C., Radeloff, V.C., Rondinini, C., Loyola, R., Davidson, A.D., 2017. Global priorities for conservation across multiple dimensions of mammalian diversity. Proc. Natl. Acad. Sci. 114 (29), 7641–7646. https://doi.org/10.1073/pnas.1706461114.

Cox, C., Ladle, R., Moore, P., 2020. Biogeography: An Ecological and Evolutionary Approach, 10th ed. Wiley.

Dansereau, P., 1957. Biogeography: An Ecological Perspective. Ronald Press Co.

Dolédec, S., Chessel, D., Gimaret-Carpentier, C., 2000. Niche separation in community analysis: a new method. Ecology 81 (10), 2914–2927. https://doi.org/10.1890/0012-9658(2000)081[2914:NSICAA]2.0.CO;2.

Elith, J., Leathwick, J., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40 (1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Eronen, J., Polly, P., Fred, M., Damuth, J., Frank, D., Mosbrugger, V., Scheidegger, C., Stenseth, N., Fortelius, M., 2010. Ecometrics: the traits that bind the past and present together. Integr. Zool. 5 (2), 88–101. https://doi.org/10.1111/j.1749-4877.2010.00192.x.

Ficetola, G., Mazel, F., Thuiller, W., 2017. Global determinants of zoogeographical boundaries. Nat. Ecol. Evol. 1, 1–7. https://doi.org/10.1038/s41559-017-0089.

Fortelius, M., et al., 2002. Fossil mammals resolve regional patterns of Eurasian climate change over 20 million years. Evol. Ecol. Res. 4 (7), 1005–1016.

Galbrun, E., Miettinen, P., 2012. From black and white to full color: extending redescription mining outside the Boolean world. Stat. Anal. Data Min. 5 (4), 284–303. https://doi.org/10.1002/sam.11145.

Galbrun, E., Miettinen, P., 2017. Redescription Mining. Springer. https://doi.org/10.1007/978-3-319-72889-6.

Galbrun, E., Miettinen, P., 2018. Mining redescriptions with Siren. ACM Trans. Knowl. Discov. Data 12 (1), 6. https://doi.org/10.1145/3007212, 1–6:30.

Galbrun, E., Tang, H., Fortelius, M., Žliobaitė, I., 2018. Computational biomes: the ecometrics of large mammal teeth. Paleontol. Electron. https://doi.org/10.26879/786.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer. https://doi.org/10.1007/978-0-387-84858-7.

He, J., Kreft, H., Gao, E., Wang, Z., Jiang, H., 2017. Patterns and drivers of zoogeographical regions of terrestrial vertebrates in China. J. Biogeogr. 44, 1172–1184. https://doi.org/10.1111/jbi.12892.

He, J., Yan, C., Holyoak, M., Wan, X., Ren, G., Hou, Y., Xie, Y., Zhang, Z., 2018. Quantifying the effects of climate and anthropogenic change on regional species loss in China. PLoS ONE 13 (7), e0199735. https://doi.org/10.1371/journal.pone.0199735.

He, J., Lin, S., Li, J., Yu, J., Jiang, H., 2020. Evolutionary history of zoogeographical regions surrounding the Tibetan Plateau. Commun. Biol. 3 (1), 1–9. https://doi.org/10.1038/s42003-020-01154-2.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83 (7), 2027–2036. https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2.

Holt, B., et al., 2013. An update of Wallace's zoogeographic regions of the world. Science 339 (6115), 74–78. https://doi.org/10.1126/science.1228282.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24 (6), 417–441. https://doi.org/10.1037/h0071325.

Huang, Y., Jacques, F., Su, T., Ferguson, D., Tang, H., Chen, W., Zhou, Z., 2015. Distribution of Cenozoic plant relicts in China explained by drought in dry season. Sci. Rep. 5, 14212. https://doi.org/10.1038/srep14212.

Jaccard, P., 1901. Étude comparative de la distribution orale dans une portion des Alpes et des Jura. Bull. Soc. Vaudoise Sci. Nat. 37, 547–579.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice-Hall, Inc.

Jongman, R., ter Braak, C., van Tongeren, O., 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press.

Krapu, C., Borsuk, M., 2020. A spatial community regression approach to exploratory analysis of ecological data. Methods Ecol. Evol. 11 (5), 608–620. https://doi.org/10.1111/2041-210X.13371.

Kreft, H., Jetz, W., 2010. A framework for delineating biogeographical regions based on species distributions. J. Biogeogr. 37 (11), 2029–2053. https://doi.org/10.1111/j.1365-2699.2010.02375.x.

Kruskal, J., 1964. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29 (2), 115–129. https://doi.org/10.1007/BF02289694.

Kulczynski, S., 1928. Die Panzenassoziationen der Pieninen. Bull. Acad. Polon. Sci. ct Lettr. C 1.

Legendre, P., Legendre, L., 2012. Numerical Ecology. Elsevier.

Liu, L., Puolamäki, K., Eronen, J., Mirzaie Ataabadi, M., Hernesniemi, E., Fortelius, M., 2012. Dental functional traits of mammals resolve productivity in terrestrial ecosystems past and present. Proc. R. Soc. Lond. B Biol. Sci. 279, 2793–2799. https://doi.org/10.1098/rspb.2012.0211.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.

MacArthur, R., Wilson, E., 1967. The Theory of Island Biogeography. Princeton University Press.

Mellin, C., Mengersen, K., Bradshaw, C., Caley, M., 2014. Generalizing the use of geographical weights in biodiversity modelling. Glob. Ecol. Biogeogr. 23 (11), 1314–1323. https://doi.org/10.1111/geb.12203.

Mihelčič, M., Šimič, G., Babić-Leko, M., Lavrač, N., Džeroski, S., Šmuc, T., 2017. Using redescription mining to relate clinical and biological characteristics of cognitively impaired and Alzheimer's disease patients. PLoS ONE 12 (10), e0187364. https://doi.org/10.1371/journal.pone.0187364.

Oksanen, O., Žliobaitė, 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. J. Biogeogr. 46 (8), 1760–1776. https://doi.org/10.1111/jbi.13586.

Olson, D., Dinerstein, E., 2002. The global 200: priority ecoregions for global conservation. Ann. Mo. Bot. Gard. 89, 125–126.

Ordoñez, J., Van Bodegom, P., Witte, J.-P., Wright, I., Reich, P., Aerts, R., 2009. A global study of relationships between leaf traits, climate and soil measures of nutrient fertility. Glob. Ecol. Biogeogr. 18 (2), 137–149. https://doi.org/10.1111/j.1466-8238.2008.00441.x.

Ovaskainen, O., Abrego, N., 2020. Joint Species Distribution Modelling. Cambridge University Press.

Pearse, W., Peres-Neto, P. (Eds.), 2017. Methods in Ecology and Evolution. Special issue: Biogeography.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Lond. Edinburgh Dublin Philos. Mag. J. Sci. 2 (11), 559–572. https://doi.org/10.1080/14786440109462720.

Ramakrishnan, N., Zaki, M., 2009. Redescription mining and applications in bioinformatics. In: Biological Data Mining.

Ramakrishnan, N., Kumar, D., Mishra, B., Potts, M., Helm, R., 2004. Turning CARTwheels: An alternating algorithm for mining redescriptions. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 266–275. https://doi.org/10.1145/1014052.1014083.

Ratnam, J., Tomlinson, K.W., Rasquinha, D.N., Sankaran, M., 2016. Savannahs of Asia: antiquity, bio609 geography, and an uncertain future. Phil. Trans. R. Soc. B: Biol. Sci. 371 (1703), 20150305. https://doi.org/10.1098/rstb.2015.0305.

Tang, Z., Wang, Z., Zheng, C., Fang, J., 2006. Biodiversity in China's mountains. Front. Ecol. Environ. 4 (7), 347–352. https://doi.org/10.1890/1540-9295(2006)004[0347:BICM]2.0.CO;2.

Thomas, H., Myers-Smith, I., Bjorkman, A., Elmendorf, S., Blok, D., Cornelissen, J., et al., 2019. Traditional plant functional groups explain variation in economic but not size-related traits across the tundra biome. Glob. Ecol. Biogeogr. 28 (2), 78–95. https://doi.org/10.1111/geb.12783.

Vavrek, M., 2016. A comparison of clustering methods for biogeography with fossil datasets. PeerJ 4, e1720.

Vermillion, W., Polly, P., Head, J., Eronen, J., Lawing, A., 2018. Ecometrics: A trait-based approach to paleoclimate and paleoenvironmental reconstruction. In: Croft, Darin A., Su, Denise F., Simpson, Scott W. (Eds.), Methods in Paleoecology: Reconstructing Cenozoic Terrestrial Environments and Ecological Communities, pp. 373–394.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Yamada, K., Kohara, K., Ikehara, M., Seto, K., 2019. The variations in the east Asian summer monsoon over the past 3 kyrs and the controlling factors. Sci. Rep. 9 (1), 5036. https://doi.org/10.1038/s41598-019-41359-y.

Zhao, C., Wang, Y., Yang, Q., Fu, R., Cunnold, D., Cho, Y., 2010. Impact of east Asian summer monsoon on the air quality over China: view from space. J. Geophys. Res. 115 (D9) https://doi.org/10.1029/2009JD012745.

Žliobaitė, I., Rinne, J., Tóth, A., Mechenich, M., Liu, L., Behrensmeyer, A., Fortelius, M., 2016. Herbivore teeth predict climatic limits in Kenyan ecosystems. Proc. Natl. Acad. Sci. 113 (45), 12751–12756. https://doi.org/10.1073/pnas.1609409113.

Žliobaitė, I., Tang, H., Saarinen, J., Fortelius, M., Rinne, J., Rannikko, J., 2018. Dental ecometrics of tropical Africa: linking vegetation types and communities of large plant-eating mammals. Evol. Ecol. Res. 19, 127–147.