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

Appendix

Original article published in Ecological Informatics (2021)
https://doi.org/10.1016/j.ecoinf.2021.101314

Authors:

Esther Galbrun (corresponding author)
esther.galbrun@uef.fi
School of Computing, University of Eastern Finland,
Technopolis, Microkatu 1, FI-70210 Kuopio, Finland.

Hui Tang
Department of Geosciences, University of Oslo,
P.O. Box 1022, University of Oslo, NO-0315 Oslo, Norway.

Anu Kaakinen
Department of Geosciences and Geography, University of Helsinki,
P.O. Box 64, FI-00014 University of Helsinki, Finland.

Indrė Žliobaitė
Department of Computer Science, University of Helsinki,
P.O. Box 64, FI-00014 University of Helsinki, Finland.

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.

A Materials for the case study

A.1 Delimitation of the study region

The coverage of our dataset is shown in Fig. 1. Localities depicted in green belong to our study area covering China and southern Asia, while other localities included in the global dataset of (Galbrun et al., 2018) are depicted in grey. The study region (green) corresponds to the union of the five rectangles defined by the corners given in Table 1. The rectangles were selected because they demarcate the region we are interested in while also taking advantage of the natural boundaries and the boundaries generated by the requirement on the minimum number of distinct taxa.

Table 1: Coordinates of the corners of the five rectangles used to define the study region.

North-West	20°N, 80°E	10°N, 90°E	5°N, 66°E	28° N, 67°30′E	35°N, 80°E
South-East	35°N, 125°E	20°N, 115°E	28°N, 90°E	37°30′N, 90° E	40N, 120°E

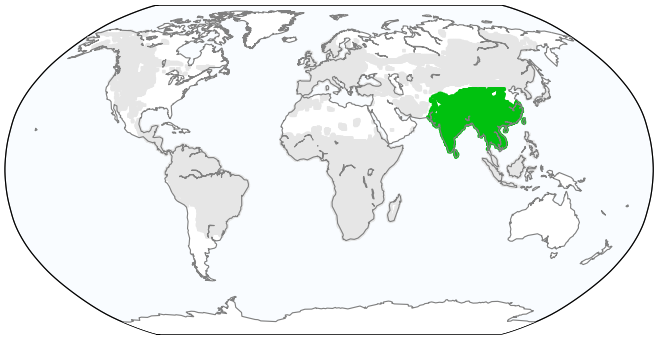


Fig. 1: Maps of the localities contained in the dataset. Localities depicted in green belong to our study area covering China and southern Asia, while other localities included in the global dataset of (Galbrun et al., 2018) are depicted in grey.

A.2 Datasets

Fig. 2 shows the data aggregation process. Fig. 3, 4 and 5 show the distribution of dental and climatic variables within the study region.

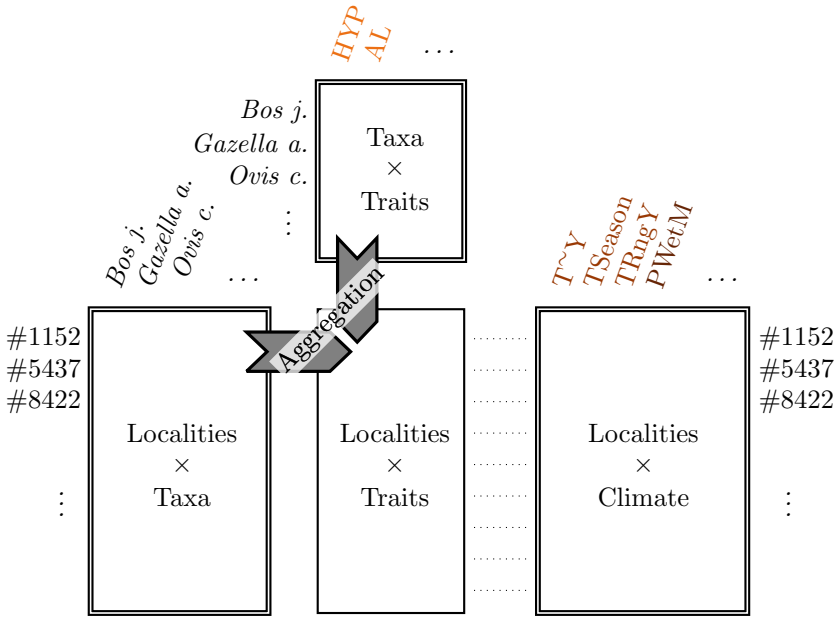


Fig. 2: Datasets and data aggregation. The initial datasets (Localities × Taxa) and (Taxa × Traits) are aggregated to produce the (Localities × Traits) dataset.

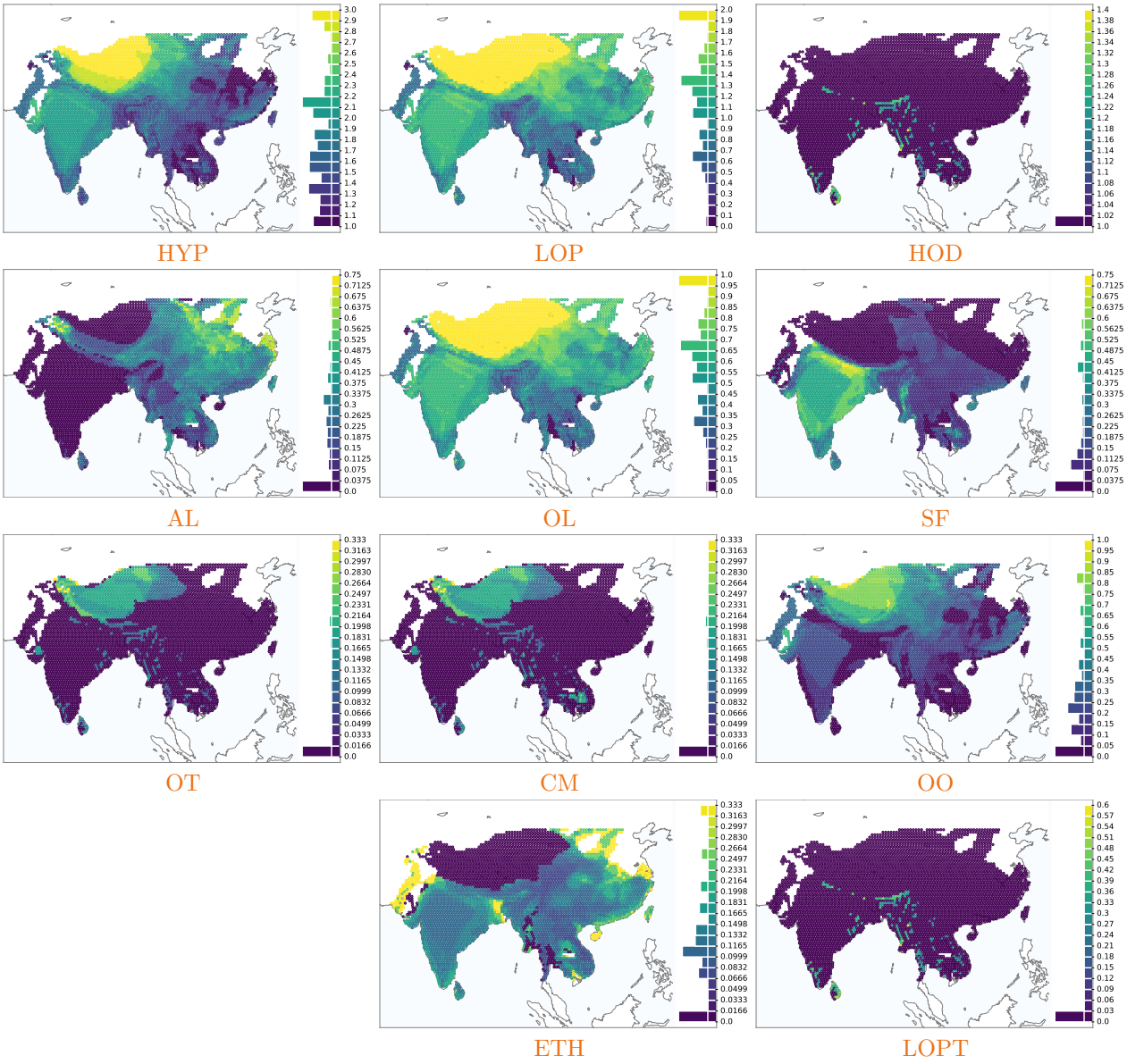


Fig. 3: Maps of dental trait distributions. The color of each dot denotes the value of the trait variable at the corresponding locality. The colored horizontal histogram on the right provides the legends for the colors and indicates the frequency of the different values. The plots are at high resolution intended for electronic viewing and can be zoomed in.

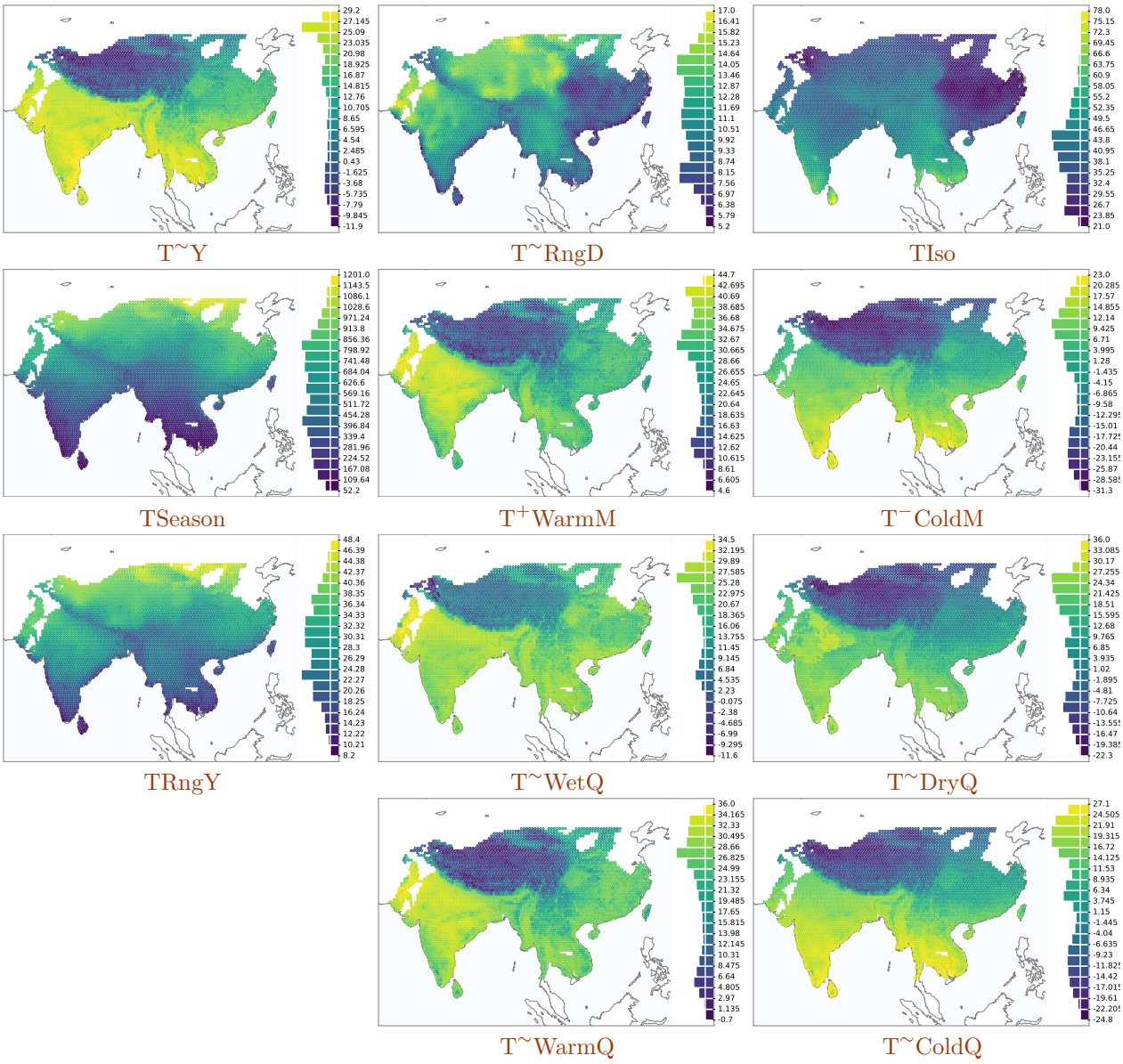


Fig. 4: Maps of bioclimatic variables: temperatures. The color of each dot denotes the value of the temperature variable at the corresponding locality. The colored horizontal histogram on the right provides the legends for the colors and indicates the frequency of the different values. The plots are at high resolution intended for electronic viewing and can be zoomed in.

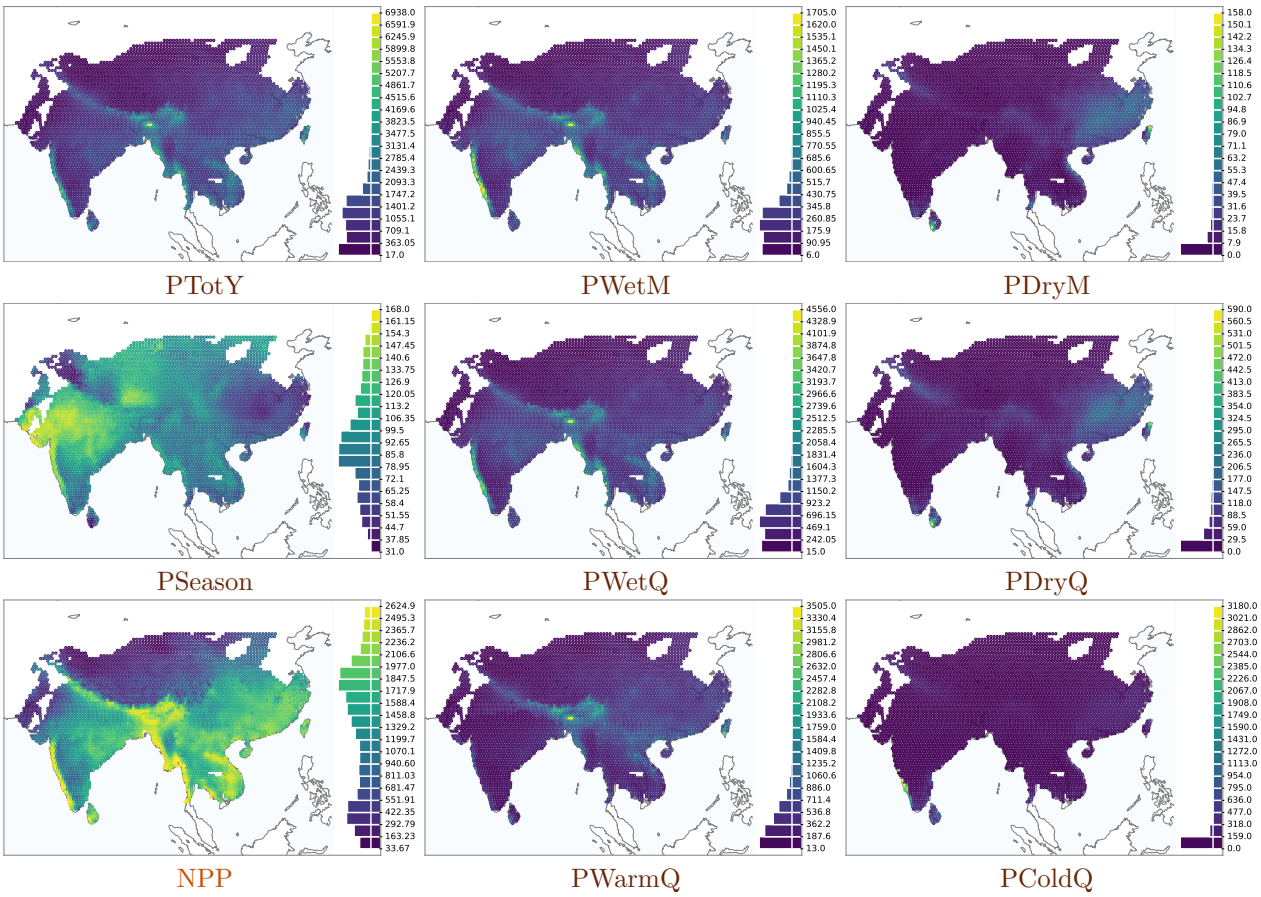


Fig. 5: Maps of bioclimatic variables: precipitation and net primary product (NPP, computed from mean annual temperature and precipitation). The color of each dot denotes the value of the variable at the corresponding locality. The colored horizontal histogram on the right provides the legends for the colors and indicates the frequency of the different values. The plots are at high resolution intended for electronic viewing and can be zoomed in.

B Preamble: classical analysis methods

B.1 Pairwise correlation and scatter plots

Pearson correlation coefficient and scatter plots for a subset of climate variables are shown in Fig. 6.

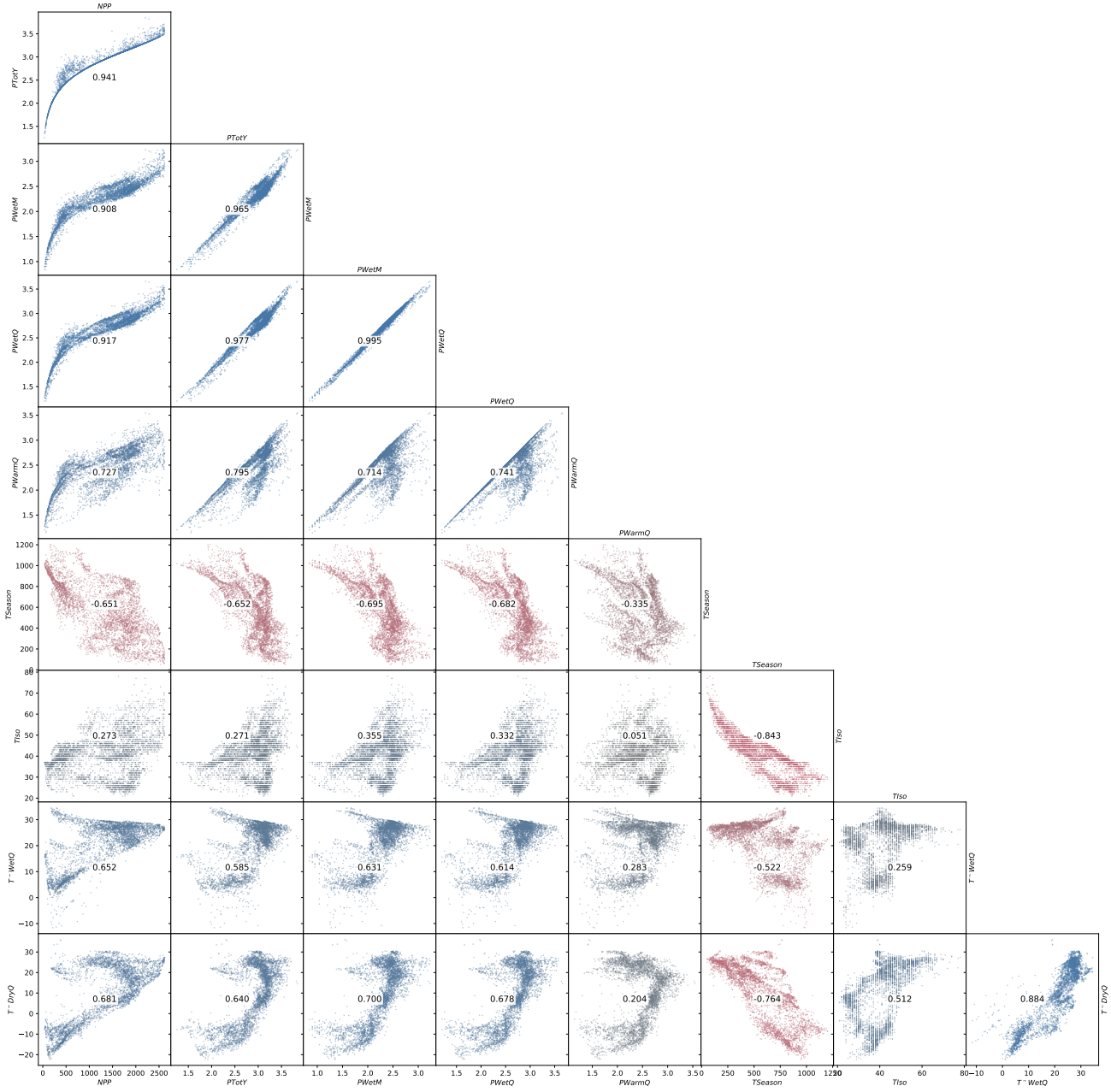


Fig. 6: Pearson correlation coefficient (r) and scatter plots for a subset of climate variables. In each plot, the dots represent the localities and are drawn in color according to the Pearson correlation coefficient of the corresponding pair of variables, from blue for strong positive correlation (r close to 1) to red for strong negative correlation (r close to -1).

B.2 Multivariate projections

PCA projections considering *Dental traits* and *Climate* variables separately and together are shown in in Fig. 7.

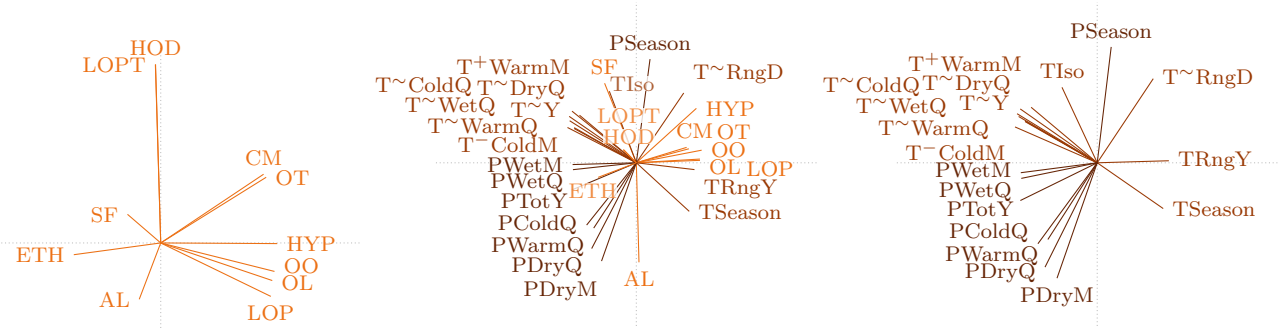


Fig. 7: PCA projections. The principal component analysis is performed and the variables are projected on the first two components considering only dental traits variables (left), only climate variables (right) and all together (middle).

Table 2: Regression models. Estimating one climate variable as a linear function of one or more dental trait variable(s) using the generalized least squares (GLS). Best-fitting single-variate, two-variate and three-variate models according to Akaike's information criterion (AIC) for three example climatic variables. For each model, we indicate the corresponding equation as well as Akaike's information criterion (AIC), the F-statistic (F), the Log-Likelihood (LL), and the mean squared error of the residuals (MSE) of the model.

model	AIC	F	LL	MSE
PTotY				
$3.63 - 0.61\,\mathrm{LOP}$	1645	4432	−820	0.08
$3.91 - 0.56\,\mathrm{HYP}$	1016	5785	−506	0.07
$3.22 - 1.35\,\mathrm{OO}$	563	6888	−279	0.07
$3.40 - 1.00\,\mathrm{OO} - 0.22\,\mathrm{LOP}$	268	3835	−131	0.06
$3.38 - 1.54\,\mathrm{OO} - 0.94\,\mathrm{ETH}$	252	3857	−123	0.06
$3.80 - 0.57\,\mathrm{HYP} + 0.71\,\mathrm{SF}$	−266	4613	136	0.06
$3.98 - 0.77\,\mathrm{HYP} + 0.53\,\mathrm{OO} + 1.01\,\mathrm{SF}$	−336	3149	172	0.05
TIso				
$50.85 - 16.25\,\mathrm{OL}$	31849	821	−15922	79.36
$50.94 - 8.26\,\mathrm{LOP}$	31790	891	−15893	78.31
$45.24 - 23.62\,\mathrm{AL}$	31564	1170	−15780	74.40
$42.16 - 43.60\,\mathrm{OL} + 14.09\,\mathrm{HYP}$	30680	1205	−15337	60.89
$54.56 - 22.41\,\mathrm{AL} - 7.72\,\mathrm{LOP}$	30662	1218	−15328	60.65
$41.77 - 21.46\,\mathrm{LOP} + 13.97\,\mathrm{HYP}$	30569	1292	−15281	59.38
$46.40 - 9.15\,\mathrm{AL} + 9.17\,\mathrm{HYP} - 16.71\,\mathrm{LOP}$	30518	889	−15255	58.68
$63.17 - 19.93\,\mathrm{AL} - 38.78\,\mathrm{ETH} - 22.78\,\mathrm{OL}$	30073	1139	−15032	53.06
$62.93 - 19.70\,\mathrm{AL} - 38.59\,\mathrm{ETH} - 11.33\,\mathrm{LOP}$	30008	1178	−15000	52.28
T~Y				
$38.56 - 36.95\,\mathrm{OL}$	30143	6250	−15069	53.93
$38.04 - 18.19\,\mathrm{LOP}$	30123	6299	−15059	53.68
$24.98 - 37.06\,\mathrm{OO}$	29942	6748	−14969	51.53
$30.84 - 40.72\,\mathrm{OO} - 25.78\,\mathrm{AL}$	27375	7776	−13684	28.81
$31.03 - 16.52\,\mathrm{LOP} + 29.48\,\mathrm{SF}$	26306	10511	−13150	22.61
$31.52 - 33.63\,\mathrm{OL} + 29.63\,\mathrm{SF}$	26288	10562	−13141	22.52
$30.53 - 15.11\,\mathrm{LOP} - 3.88\,\mathrm{OO} + 27.90\,\mathrm{SF}$	26259	7101	−13125	22.37
$30.99 - 30.81\,\mathrm{OL} - 3.81\,\mathrm{OO} + 28.06\,\mathrm{SF}$	26242	7132	−13117	22.29

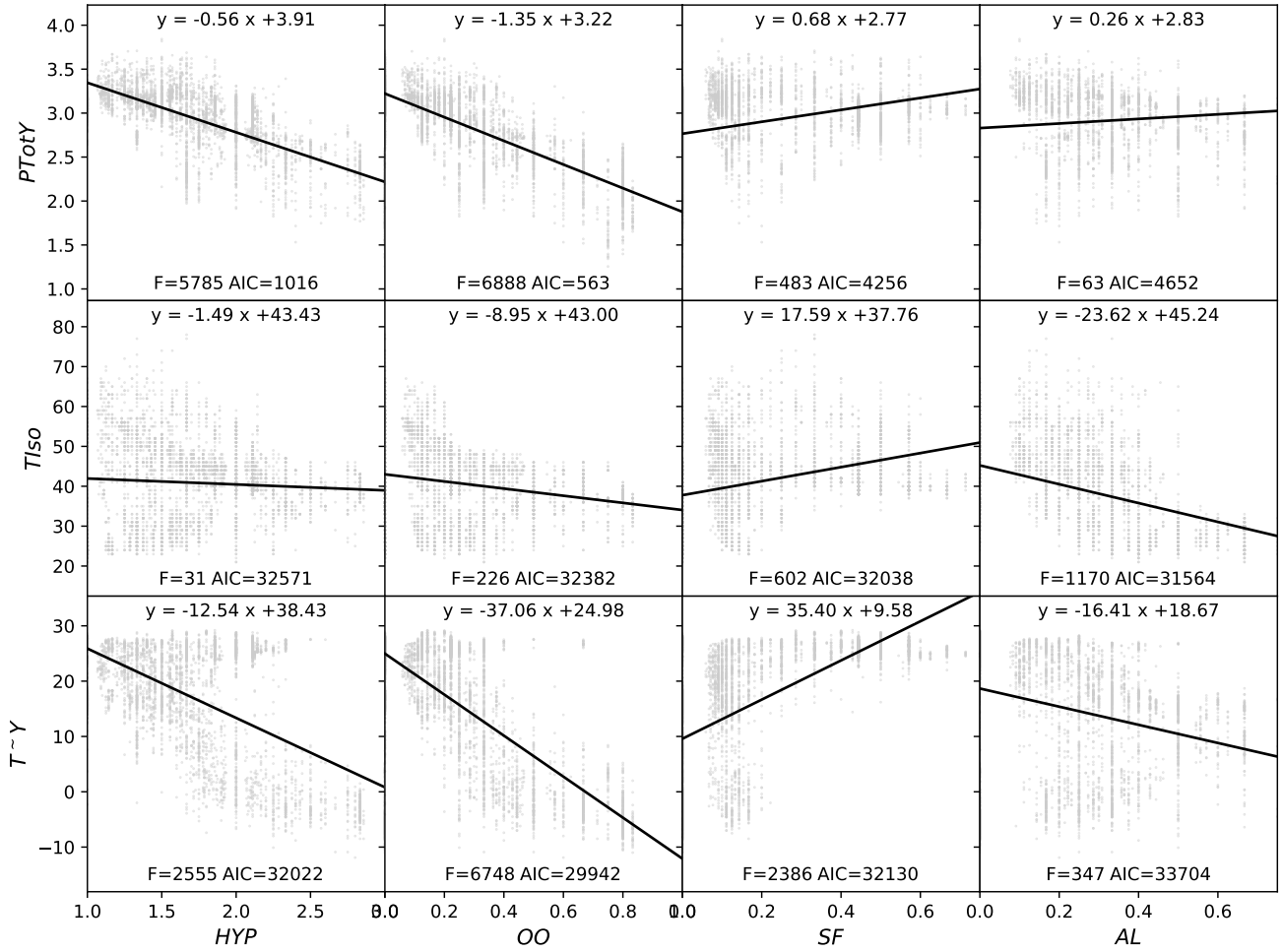


Fig. 8: Regression plots. Estimating one climate variable (y) as a linear function of a dental trait variable (x) using the generalized least squares (GLS) (Aitken, 1934). For each model, we indicate the corresponding equation (top) as well Akaike's information criterion (AIC) and the F-statistic (F) of the model (bottom).

B.3 Regression models

The best-fitting linear regression models according to Akaike's information criterion (AIC) for PTotY, TIso and T~Y are listed as examples in Table 2. Examples of single-variate models are plotted in Fig. 8.

B.4 Clustering

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:

`single` (SL), the minimum distance between cluster members, i.e. $\mathrm{D}(U, V) = \min_{(u,v)\in U\times V} \mathrm{d}(u, v)$

`complete` (CL), the maximum distance between cluster members, i.e. $\mathrm{D}(U, V) = \max_{(u,v)\in U\times V} \mathrm{d}(u, v)$

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

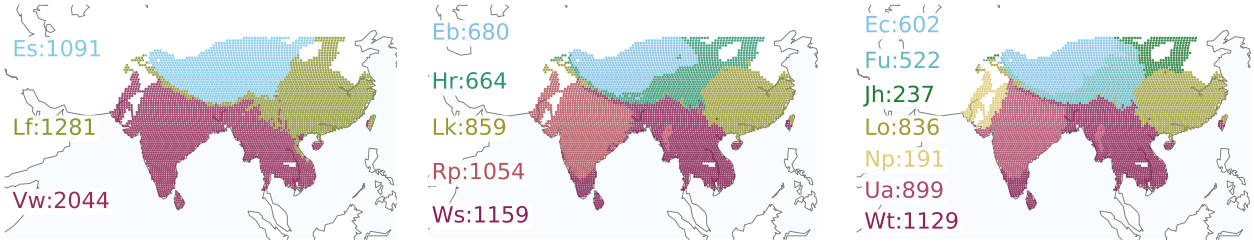
`weighted` a.k.a. Weighted Pair-Group Method using arithmetic Averages (WPGMA), a variant of UPGMA weighted by the size of the clusters

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

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

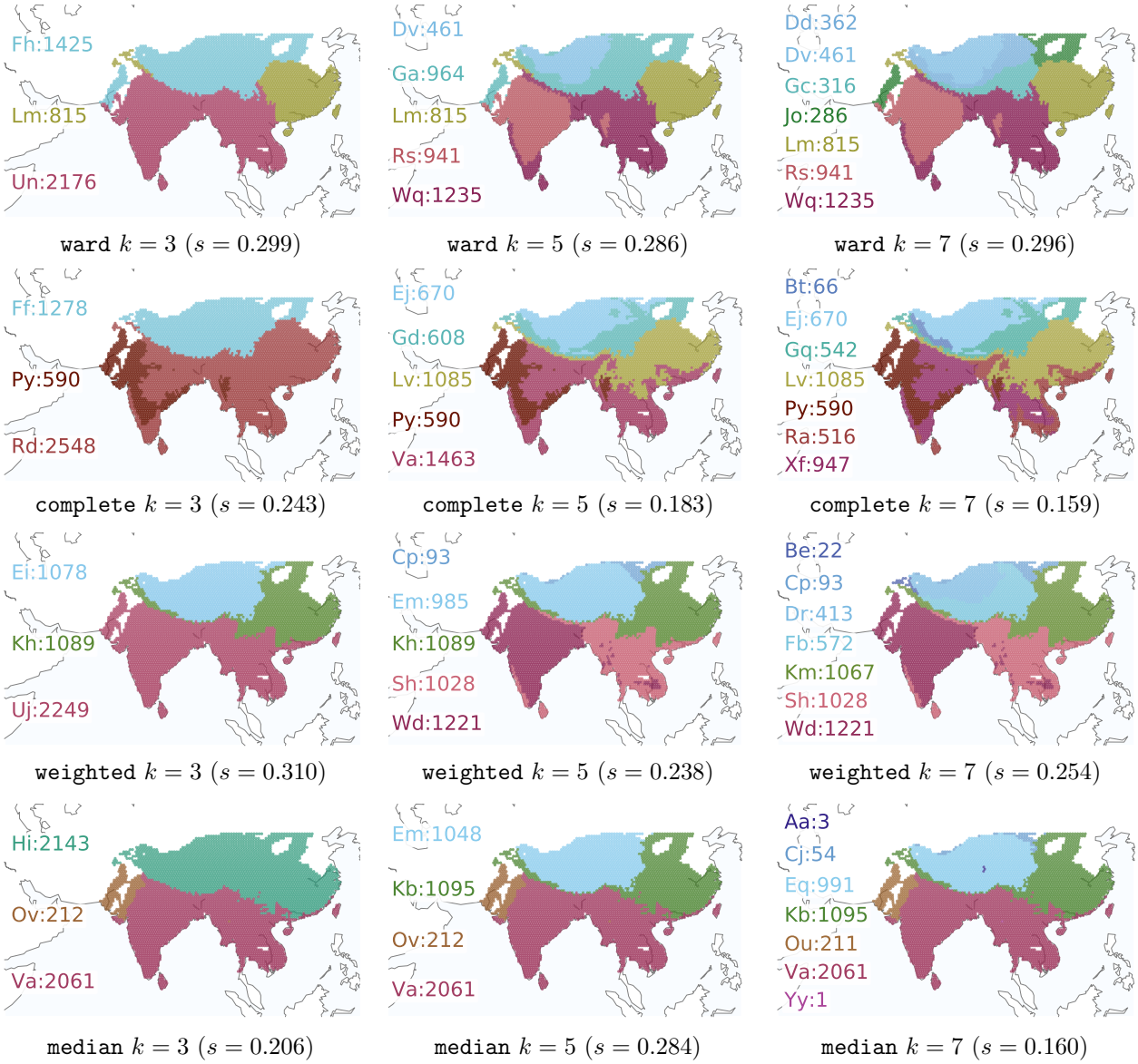
`ward` relies on Ward's minimum variance criterion

The *silhouette coefficient* was introduced by (Rousseeuw, 1987) to evaluate the quality of a clustering. We consider a set of data points E, in our case localities, and a clustering $\{C_1, C_2, \dots\}$ such that each data point



`kmeans` $k = 3$ $(s = 0.321)$ `kmeans` $k = 5$ $(s = 0.327)$ `kmeans` $k = 7$ $(s = 0.325)$

Fig. 9: k-means clusterings and associated silhouette coefficients (s). To the left of each map, we list the different clusters, with the number of localitites they contain. Labels and colors are assigned to clusters in such a way that more similar clusters (across all generated clusterings) receive similar colors and labels that are close in the alphabetical order.



`ward` $k = 3$ $(s = 0.299)$ `ward` $k = 5$ $(s = 0.286)$ `ward` $k = 7$ $(s = 0.296)$

`complete` $k = 3$ $(s = 0.243)$ `complete` $k = 5$ $(s = 0.183)$ `complete` $k = 7$ $(s = 0.159)$

`weighted` $k = 3$ $(s = 0.310)$ `weighted` $k = 5$ $(s = 0.238)$ `weighted` $k = 7$ $(s = 0.254)$

`median` $k = 3$ $(s = 0.206)$ `median` $k = 5$ $(s = 0.284)$ `median` $k = 7$ $(s = 0.160)$

Fig. 10: Hierarchical agglomerative clusterings and associated silhouette coefficients (s). Each panel corresponds to a particular combination of linkage function and number of clusters (k) as indicated underneath. To the left of each map, we list the different clusters, with the number of localitites they contain. Labels and colors are assigned to clusters in such a way that more similar clusters (across all generated clusterings) receive similar colors and labels that are close in the alphabetical order.

belongs to one cluster. Furthermore, we consider a function d that measures the distance between any pair of data points with respect to data variables.

For data point x belonging to cluster C_i, let

$$a(x) = \frac{1}{|C_i| - 1} \sum_{y \in C_i, x \neq y} \mathrm{d}(x, y) \qquad \text{and} \qquad b(x) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} \mathrm{d}(i, j) \ .$$

That is, $a(x)$ denotes the average distance between x and other points in the same cluster whereas $b(x)$ is the minimum, over the other clusters, of average distance between x and points in that cluster.

Then, the silhouette value for point $x \in C_i$ is

$$s(x) = \begin{cases} \frac{b(x) - a(x)}{\max(a(x), b(x))} & \text{if } |C_i| > 1 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

and the silhouette coefficient is the average silhouette value over all data points

$$s = \sum_{x \in E} \frac{s(x)}{|E|} \ .$$

Clusterings obtained with the k-means algorithm and with hierarchical agglomerative algorithms for $k = 3$, 5 and 7 clusters can be found in Fig. 9 and 10, respectively.

Note that the merging process in hierarchical agglomerative algorithms may result in fewer clusters than specified, as is the case here with the `median` variant where four clusters are returned when setting $k = 5$. Single linkage as well as average linkage and centroid linkage (i.e. unweighted variants) returned partitions that typically consisted of one very large cluster together with extremely small clusters (e.g. a dozen localities each), which is not very informative for further analysis. Obtaining a well-balanced hierarchical clustering is often methodologically challenging, and here it may be further complicated by the presence of spatial autocorrelation in the data.

To facilitate the visual comparison of the (many) clusters produced by the different methods, we compute a one-dimensional distance-preserving projection of the clusters (by building the graph of pair-wise similarities between the clusters and computing the vector associated to the second-smallest eigenvalue of the associated Laplacian matrix, a.k.a. Fiedler vector, commonly used to partition graphs)and assign colors along a gradient, in such a way that similar clusters are close to each other and assigned similar colors.

C Redescription mining methodology

To assess the statistical significance of a redescription, we compute a p-value 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. Consider two statistically independent random queries with marginal probabilities equal to those of the queries under consideration, i.e. with marginal probabilities equal to the fraction of covered localities $P(q) = |\mathrm{supp}(q)| \,/\, |E|$. The p-value is computed using the binomial distribution, as

$$\mathrm{pval}(q_{\mathbf{D}}, q_{\mathbf{C}}) = \sum_{s=|S|}^{|E|} \binom{|E|}{s} \big(P(q_{\mathbf{D}})P(q_{\mathbf{C}})\big)^s \big(1 - P(q_{\mathbf{D}})P(q_{\mathbf{C}})\big)^{|E|-s} \ ,$$

where $S = \mathrm{supp}((q_{\mathbf{D}}, q_{\mathbf{C}}))$ is the observed support of the redescription. The p-value is the probability of obtaining a set of same size or larger if each element of a set of size $|E|$ has a probability equal to the product of marginals $P(q_{\mathbf{D}})$ and $P(q_{\mathbf{C}})$ to be selected, in accordance with the independence assumption.

D Case study: biogeographic analysis with redescription mining

For reference, the top ten redescriptions from the two runs are listed in Table 3.

Summaries obtained with for the different clustering variants and for $k = 3$, 5 and 7 clusters are shown in Fig. 11. Since they are based on the support of redescriptions, the clusterings naturally take into account both dental traits and bioclimatic variables, whereas when clustering the raw data, the different nature and scale of the variables might raise issues.

Table 3: Top-ten redescriptions from the two runs. For each redescription, we list its queries, that is, the query over dental traits variables ($q_{\mathbf{D}}$) and the query over bioclimatic variables ($q_{\mathbf{C}}$). We also indicate the accuracy of the redescription (J) as well as the size of its support, as the number of localities described ($|\,\text{supp}\,|$) and as a percentage of the total number of localities (supp %).

R1.1 J = 0.846 |supp| = 2540 supp % = 57.52
$q_{\mathbf{D}}$ = [LOP ≤ 1.556] AND [OO ≤ 0.25]
$q_{\mathbf{C}}$ = [19.9 ≤ T~WetQ ≤ 29.6] AND [388 ≤ PTotY]

R1.2 J = 0.845 |supp| = 893 supp % = 20.22
$q_{\mathbf{D}}$ = [0.769 ≤ OL] AND [SF ≤ 0.222] AND [ETH ≤ 0.125]
$q_{\mathbf{C}}$ = [T~Y ≤ 5.9] AND [592.8 ≤ TSeason ≤ 1064]

R1.3 J = 0.810 |supp| = 2292 supp % = 51.90
$q_{\mathbf{D}}$ = [0.222 ≤ LOP ≤ 1.333] AND [AL ≤ 0.417]
$q_{\mathbf{C}}$ = [7.6 ≤ T~DryQ] AND [6.5 ≤ T~ColdQ ≤ 25.2]

R1.4 J = 0.810 |supp| = 2271 supp % = 51.43
$q_{\mathbf{D}}$ = [HYP ≤ 1.889] AND [LOP ≤ 1.75] AND [OO ≤ 0.4]
$q_{\mathbf{C}}$ = [19.6 ≤ T+WarmM ≤ 38.5] AND [116 ≤ PWarmQ]

R1.5 J = 0.787 |supp| = 1982 supp % = 44.88
$q_{\mathbf{D}}$ = [0.222 ≤ LOP ≤ 1.833] AND [0.077 ≤ AL ≤ 0.667] AND [OT ≤ 0.111]
$q_{\mathbf{C}}$ = [3.6 ≤ T~WarmQ ≤ 29.4] AND [266 ≤ PWarmQ ≤ 1556]

R1.6 J = 0.769 |supp| = 2025 supp % = 45.86
$q_{\mathbf{D}}$ = [HYP ≤ 2.125] AND [LOP ≤ 1.5] AND [0.059 ≤ SF ≤ 0.571]
$q_{\mathbf{C}}$ = [TRngY ≤ 30.3] AND [−1.8 ≤ T~ColdQ ≤ 26.6]

R1.7 J = 0.751 |supp| = 1959 supp % = 44.36
$q_{\mathbf{D}}$ = [AL ≤ 0.417] AND [0.059 ≤ SF] AND [0.059 ≤ ETH ≤ 0.167]
$q_{\mathbf{C}}$ = [92.7 ≤ TSeason ≤ 646] AND [631.57 ≤ NPP ≤ 2608.99]

R1.8 J = 0.749 |supp| = 2210 supp % = 50.05
$q_{\mathbf{D}}$ = [0.077 ≤ AL] AND [SF ≤ 0.25]
$q_{\mathbf{C}}$ = [−5.5 ≤ T~Y ≤ 23.9] AND [T+WarmM ≤ 34.8]

R1.9 J = 0.735 |supp| = 2168 supp % = 49.09
$q_{\mathbf{D}}$ = [LOP ≤ 1.556] AND [OO ≤ 0.4] AND [0.083 ≤ ETH ≤ 0.2]
$q_{\mathbf{C}}$ = [TSeason ≤ 870.3] AND [181 ≤ PWetM]

R1.10 J = 0.732 |supp| = 2045 supp % = 46.31
$q_{\mathbf{D}}$ = [HYP ≤ 2.444] AND [0.077 ≤ AL] AND [SF ≤ 0.375]
$q_{\mathbf{C}}$ = [14.3 ≤ T+WarmM ≤ 34.6] AND [1 ≤ PDryM]

R2.1 J = 0.836 |supp| = 741 supp % = 16.78
$q_{\mathbf{D}}$ = [1.8 ≤ LOP] AND [ETH ≤ 0]
$q_{\mathbf{C}}$ = [T~WarmQ ≤ 17.9] AND [PTotY ≤ 496]

R2.2 J = 0.829 |supp| = 2553 supp % = 57.81
$q_{\mathbf{D}}$ = [OO ≤ 0.25]
$q_{\mathbf{C}}$ = [19.9 ≤ T~WetQ ≤ 29.6]

R2.3 J = 0.822 |supp| = 1863 supp % = 42.19
$q_{\mathbf{D}}$ = [LOP ≤ 0.769] OR [0.167 ≤ SF]
$q_{\mathbf{C}}$ = [16.8 ≤ T~DryQ]

R2.4 J = 0.808 |supp| = 2494 supp % = 56.48
$q_{\mathbf{D}}$ = [0.222 ≤ LOP ≤ 1.333]
$q_{\mathbf{C}}$ = [15 ≤ T~Y ≤ 27.4]

R2.5 J = 0.783 |supp| = 2315 supp % = 52.42
$q_{\mathbf{D}}$ = [HYP ≤ 1.889]
$q_{\mathbf{C}}$ = [19.6 ≤ T+WarmM ≤ 38.5]

R2.6 J = 0.749 |supp| = 2306 supp % = 52.22
$q_{\mathbf{D}}$ = [0.077 ≤ AL]
$q_{\mathbf{C}}$ = [6.2 ≤ T~WarmQ ≤ 28.5]

R2.7 J = 0.717 |supp| = 2212 supp % = 50.09
$q_{\mathbf{D}}$ = [0.111 ≤ ETH]
$q_{\mathbf{C}}$ = [29.1 ≤ T+WarmM]

R2.8 J = 0.708 |supp| = 2185 supp % = 49.48
$q_{\mathbf{D}}$ = [0.083 ≤ ETH ≤ 0.2]
$q_{\mathbf{C}}$ = [181 ≤ PWetM]

R2.9 J = 0.685 |supp| = 2109 supp % = 47.76
$q_{\mathbf{D}}$ = [0.059 ≤ SF ≤ 0.571]
$q_{\mathbf{C}}$ = [TRngY ≤ 30.3]

R2.10 J = 0.681 |supp| = 2039 supp % = 46.17
$q_{\mathbf{D}}$ = [LOP ≤ 1.286]
$q_{\mathbf{C}}$ = [T~RngD ≤ 12.3]

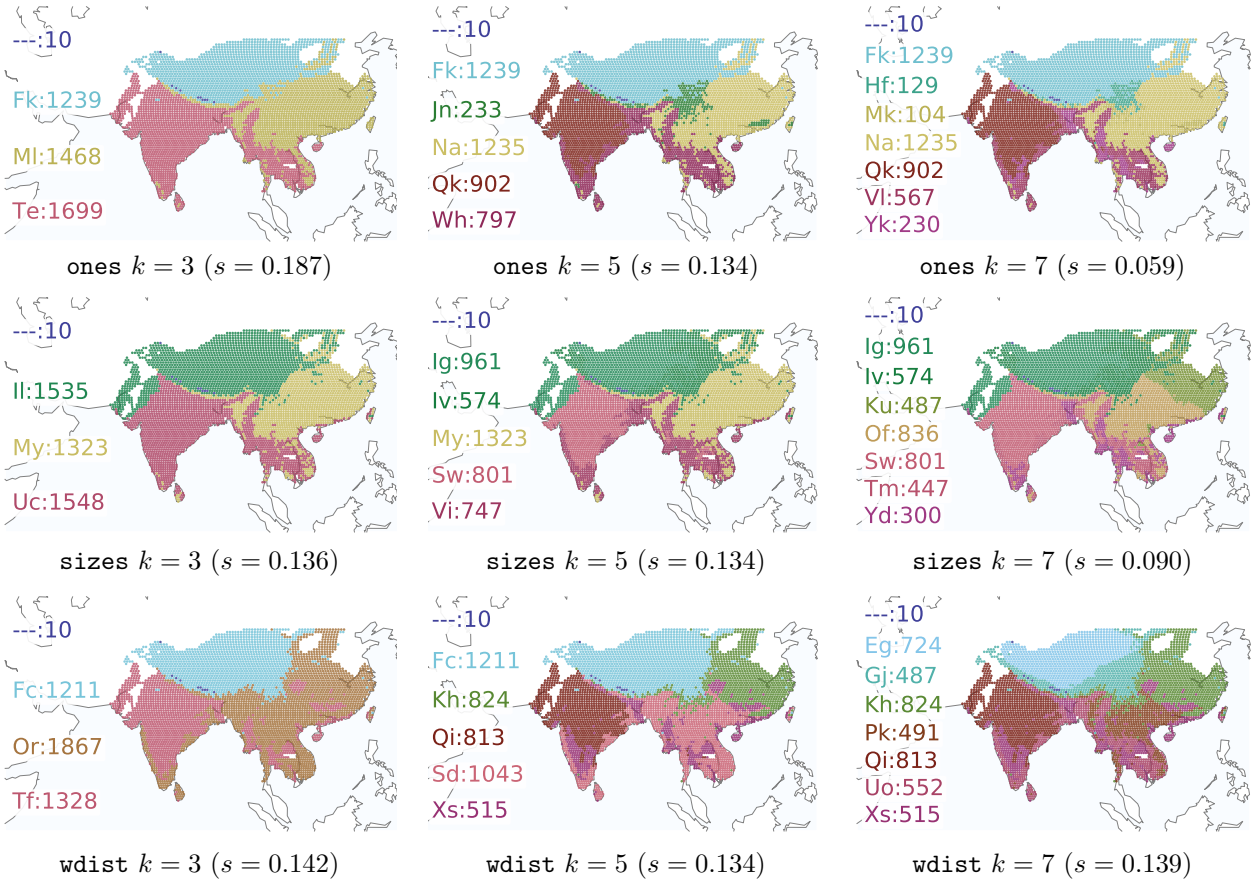


Fig. 11: Redescription summaries, i.e. redescription-based clusterings, and associated silhouette coefficients (s). To the left of each map, we list the different clusters, with the count of localitites they contain. Labels and colors are assigned to clusters in such a way that more similar clusters (across all generated clusterings) receive similar colors and labels that are close in the alphabetical order. There are a few localities (in dark blue) that do not support any of the selected redescriptions and hence do not belong to any cluster.

Compared to the results of the clustering analysis based on raw dental traits and climate data (cf. Fig. 9 and 10), the redescription summaries (cf. Fig. 11) reveal generally similar distinct biogeographical regions, 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, especially when considering few clusters (i.e. for $k = 3$ or 5). With more clusters ($k = 5$ or 7), the results based on redescriptions reveal a much finer spatial structure over southern China and Southeast Asia, which corresponds well to plant relicts found in these regions Huang et al. (2015). In contrast, the results based on raw data seem to focalize on the differences within the Tibetan Plateau. These discrepancies exemplify the potential added value of using redescriptions-based clusters to delineate biologically meaningful ecoregions.

References

Aitken, A. C. (1934). "On least squares and linear combination of observations". In: *Proceedings of the Royal Society of Edinburgh*. Section B: Biological Sciences 55, pp. 42–48.

Galbrun, E., H. Tang, M. Fortelius, and I. Žliobaitė (2018). "Computational biomes: The ecometrics of large mammal teeth". In: *Paleontologia Electronica*. DOI: 10.26879/786.

Huang, Y., F. Jacques, T. Su, D. Ferguson, H. Tang, W. Chen, and Z. Zhou (2015). "Distribution of Cenozoic plant relicts in China explained by drought in dry season". In: *Scientific Reports* 5, p. 14212. DOI: 10.1038/srep14212.

Rousseeuw, P. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.