Received: 30 September 2019 | Revised: 7 November 2019 | Accepted: 7 December 2019

DOI: 10.1002/rcm.8692

RESEARCH ARTICLE

Rapid Communications in Mass Spectrometry WILEY

Which isotopes should we choose? Entropy-based feature ranking enables evaluation of the information content of stable isotopes in archaeofaunal material

Andrea Göhring[1] | Markus Mauder[2] | Peer Kröger[2] | Gisela Grupe[1]

[1] Faculty of Biology, Department of Biology I, Anthropology and Human Genomics, Ludwig Maximilian University Munich, Planegg-Martinsried, Germany

[2] Faculty of Mathematics, Computer Science and Statistics, Department of Computer Science, Database Systems Group, Ludwig Maximilian University Munich, Munich, Germany

Correspondence
A. Göhring, Ludwig Maximilian University Munich, Faculty of Biology, Department of Biology I, Anthropology and Human Genomics, Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany.
Email: andrea.goehring@lrz.uni-muenchen.de

Funding information
Archaeobiocenter of the Ludwig Maximilian University, Munich; German Research Foundation (DFG), Grant/Award Number: Gr959/15-1,2

Rationale: Methods for multi-isotope analyses are gaining in importance in anthropological, archaeological, and ecological studies. However, when material is limited (i.e., archaeological remains), it is obligatory to decide *a priori* which isotopic system(s) could be omitted without losing information.

Methods: We introduce a method that enables feature ranking of isotopic systems on the basis of distance-based entropy. The feature ranking method is evaluated using Gaussian Mixture Model (GMM) clustering as well as a cluster validation index ("trace index").

Results: Combinations of features resulting in high entropy values are less important than those resulting in low entropy values structuring the dataset into more distinct clusters. Therefore, this method allows us to rank isotopic systems. The isotope ranking depends on the analyzed dataset, for example, consisting of terrestrial mammals or fish. The feature ranking results were verified by cluster analysis.

Conclusions: Entropy-based feature ranking can be used to *a priori* select the isotopic systems that should be analyzed. Consequently, we strongly suggest that this method should be applied if only limited material is available.

1 | INTRODUCTION

1.1 | Stable isotopes

Stable isotopes have been used for anthropological and archaeological studies for many years. They contain a variety of information about, for example, the diet of ancient populations, migration patterns, environmental conditions, and climate. Skeletal remains constitute the major research substrate in physical anthropology and archaeozoology. Every skeleton is unique and, under some circumstances (e.g., exceptional burial contexts such as mummies or *pars pro toto* burials, where only pieces of a skeleton are exemplarily buried, very old archaeological age, i.e. fossils), sampling is extremely limited.

Some stable isotopic systems are linked to each other, consequently leading to consistent results. Dietary habits can be reconstructed by bone $\delta^{13}C_{collagen}$, $\delta^{13}C_{carbonate}$, and $\delta^{15}N_{collagen}$ values.[1,2] All these isotopic systems should show similar results when explaining diet. Although these isotopic systems are not identical, their output can be interpreted in a similar way.

Diet can also partly be assessed by oxygen stable isotope ratios due to water incorporated with food.[1,2] Therefore, there is an additional link of different isotopic systems. On the contrary, differences in diet can also be explained by a different origin or even a different ecological niche of an individual. This signal is also contained in oxygen stable isotope ratios.

Furthermore, the oxygen isotope ratios of bone carbonate and phosphate are partly linked. They have similar sources; however,

$\delta^{18}O_{phosphate}$ values are more influenced by the isotopic composition of drinking water, while $\delta^{18}O_{carbonate}$ values rather depend on diet.[1] Nevertheless, both can give hints on water source, diet, and a variety of environmental information, for example, climate, altitude, and latitude.[3]

Due to these relationships the question arises whether it is mandatory to measure all isotopic systems.

Multi-dimensional isotope analyses using modern data mining methods are capable of interpreting isotopic data in a more detailed way than common bivariate analyses.[4-7] For example, cluster analysis of isotopic data of fish from Haithabu and Schleswig using Gaussian Mixture Model (GMM) clustering (see 1.3 and 2.3) not only separated fish according to their habitat (freshwater, brackish, marine) but also revealed a fourth cluster of probably non-local fish from a colder environment. These groups could not be detected in the bivariate plots.[4] Therefore, it is advisable to use multi-dimensional isotopic data whenever possible. However, especially in the case of archaeological material, it is not always possible to measure all different isotopic systems due to a lack of sample material. Furthermore, some isotopic systems might not be useful for answering the research question anyway. Consequently, it is necessary to select certain isotopic systems with respect to a basic hypothesis that should be tested. However, which isotopes could be omitted without too much loss in information for the dataset? Which isotopes should we choose? In this paper, we present entropy-based feature ranking of multi-isotope fingerprints established on archaeozoological finds from a particular complex ecosystem (see section 2.1).

Therefore, the aim of this study was to evaluate isotopic systems of different subsets in order to examine if there were certain isotopic systems that do not contain much (additional) information in general and are consequently not necessarily be used if (archaeological) material is insufficient. Furthermore, combinations of certain isotopes may contain more information than others. We tested several subsets with different potential research questions (e.g., diet and non-local origin) in order to analyze which combinations of isotopes contained the highest information and which isotopic systems could be omitted due to only low information content.

A distance-based entropy measure (see below) was used to rank the isotopic systems as well as the different combinations of these systems to evaluate how important (in terms of information content) the different stable isotopes were in general with respect to the basic research questions. These research questions might include, for example, the detection of primarily non-local individuals, individuals of different cultural or social status and thus also individuals with different dietary habits, and individuals of different habitats and different ecosystems. Consequently, we expect that the dataset is separated into different groups, i.e. clusters, if the isotopic systems analyzed exhibit some information content. A clustered data structure is an important prerequisite for the method described below.

In addition, features, which are measurable properties or characteristics of e.g. an individual, are often correlated to each other. This might have an impact on the feature ranking results. Therefore, we investigated the impact of both marginal and partial correlations on feature ranking (see 2.4).

In the following, a distance-based entropy measure is introduced, which allows to differentiation between datasets with and datasets lacking any clustering structure without actually performing cluster analysis. This can be used to rank different features (here: isotopic systems) of a dataset according to their information content.

1.2 | Feature ranking using entropy

The aim of feature ranking is to find the most important feature for a specific task. A variety of methods are available for feature ranking. Recently, the Adjusted Rand Index (ARI) has been applied to multi-isotope data to test the relative contribution and importance of $\delta^{18}O_{phosphate}$ values for provenance analysis.[7-9] However, feature ranking can also be based on entropy. Entropy is a measure of the information, choice, and uncertainty of a certain variable.[10] The entropy value of a variable corresponds to its information content.[11] Shannon entropy H is defined as

$$H = -K\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

with constant K ($K > 0$) and probabilities p_i.[10]

However, in the present study we refer to a modified definition of entropy, namely a distance-based entropy measure. The probability of points, which is needed for Shannon's entropy (see Equation (1)), is usually not known. Therefore, a proxy method was applied to estimate the entropy. Accordingly, distances between data points instead of probabilities are used.[12] Entropy is not necessarily a probabilistic measure as in the basic definition by Shannon.[10] A common data mining approach is to choose a distance-based entropy as a measure of information.[12-16]

Distance-based entropy H_d can be expressed as

$$H_d = \sum_{X_i}\sum_{X_j} -D_{ij}\log_2 D_{ij} \tag{2}$$

with normalized distance matrices D_{ij} between instances X_i and X_j (see section 2.2 for more details). This entropy measure allows distinguishing between a dataset with clusters and another dataset missing any clustering structure. If the dataset is not structured into clusters, entropy is much higher than in a clearly clustered dataset. This can be explained by the fact that in a dataset containing some clusters intra-cluster distances are smaller than inter-cluster distances, resulting in lower overall distance values. Minimum distance–based entropy should thus define the optimal combination of features.[12] Therefore, it can be used for feature ranking. It is important to mention that Equation (2) does not imply any equilibration of probabilities (see Equation (1)) and distances. However, this proxy method still results in an entropy measure that can distinguish between unstructured and well-structured datasets.[12]

1.3 | Gaussian Mixture Model clustering

For an illustration of the feature ranking results, optimal combinations of features were visualized using Gaussian Mixture Model (GMM) clustering (see 2.3). Cluster analysis based on GMM clustering has already been tested for multi-isotope data.[4,6] GMM clustering is a clustering method representing data as a mixture of multivariate normal (Gaussian) distributions. The GMM clustering procedure used in the present study (R package "mclust"[17]) uses an expectation maximization (EM) algorithm to detect the maximum likelihood of the model. For additional information on GMM clustering see Göhring et al[4] and references therein.

Clustered combinations were afterward validated according to the invariant clustering criterion ("trace index") tr ($S_W{}^{-1}S_B$) where "tr" describes the trace of a scatter matrix. This validation index measures the ratio of between-cluster scatter (S_B) to within-cluster scatter (S_W).[18,19] The trace index was used for cluster validation in the different subsets with varying combinations of isotopic systems. It increases with a higher ratio of between-cluster scatter to within-cluster scatter. Consequently, the combination of isotopic systems in each dimension exhibiting a maximum trace index was identified as the result with the highest cluster quality.[19]

2 | EXPERIMENTAL

2.1 | Sample material

To examine the feature ranking method (see section 2.2) we chose subsets of the huge isotopic dataset of animal remains, which was made up of a total of 440 individuals, recovered from the Viking Haithabu (AD 804–1066) and medieval Schleswig (AD 1070–1350) sites in northern Germany located at the Schlei inlet on the Jutland Peninsula close to the Baltic Sea.[20,21] Haithabu and its successor town Schleswig were important trade centers at their time. During the ninth century Haithabu became the leading trade center of the Danish Empire. However, after Haithabu was burned down by the northern king Harald Harðráði in 1050 and after the invasion of the Western Slaves in 1066, the settlement was moved from the southern to the northern border of the Schlei inlet, where Schleswig was built.[22,23] Schleswig's influence rapidly increased in the late 11th and early 12th centuries, and Schleswig became an important transit harbor for international East–West trade. However, several economic and political factors finally led to a decline in the power of Schleswig in the late 13th century.[24,25] Both sites are influenced by the brackish water environment of the Baltic Sea. Isotope analyses might be particularly complex in this environment. Therefore, this dataset is well suited for demonstrating data mining methods. Two subsets of the huge isotopic dataset of Haithabu and Schleswig were chosen for this study, including four (dataset I: $\delta^{13}C_{collagen}$, $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$, $\delta^{18}O_{carbonate}$) and five (dataset II: $\delta^{13}C_{collagen}$, $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$, $\delta^{18}O_{carbonate}$, $\delta^{18}O_{phosphate}$) isotopic systems. Dataset I includes isotopic ratios of terrestrial mammals (herbivores, carnivores, omnivores) and fish, while dataset II includes terrestrial mammals only (Table S4, supporting information S1).

The protocols for the extraction procedures of bone collagen, carbonate, and phosphate for the samples from Haithabu and Schleswig are available online as supporting information S3.

2.2 | Entropy-based feature ranking

Entropy-based feature ranking was used to quantify the information content of different features (here: isotopic systems). Our feature ranking method aims to identify feature combinations with lowest entropy measure (see below) and thus highest information content. In addition, this method can be used to identify one or more isotopic systems, which could be omitted from the dataset without losing too much information. This can be of interest if sample material is limited, i.e. in an archaeological context.

Our entropy-based feature ranking method is modified after the work by Dash et al[12] with several changes to the procedure (see below). Prior to the ranking, we added a multi-dimensional outlier detection method using the R package "mvoutlier".[26] For the outlier detection procedure robust principal components were computed from the robustly sphered and normed data in order to compute the covariance matrix. Based on this matrix Mahalanobis distances were calculated. The 97.5th quantile of the χ^2 distribution was used as an outlier cutoff value.[26,27] Outliers might strongly influence feature ranking results, especially if a distance-based method is conducted as in this study (see results). Furthermore, as isotope ratios of different systems can exhibit quite different measurement ranges, the original (non-transformed) data points were normalized prior to the ranking to eliminate this influencing factor using the following formula resulting in values between zero and one:

$$\frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

Afterward, distance matrices D_{ij} were computed using Euclidean distances

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{4}$$

between all (normalized) data points. The distance matrices created were again normalized according to Equation (3).

As mentioned earlier, the entropy measure used in this study is not identical to Shannon's definition of entropy (see Equation (1)), but uses distances instead of probabilities (see Equation (2)).

The distance-based entropy measure used in this study is similar to that of Dash et al,[12] however with three main differences, which shall be explained in the following:

First, Dash et al[12] chose the formula similar to Shannon's entropy based on two possibilities with probabilities p and q ($q = 1 - p$; see Shannon[10]). However, while working with distances instead of probabilities, we decided to focus on the calculated distance matrix (D_{ij}) instead of an additional complementary distance ($1–D_{ij}$; see Dash et al[12]), which actually has no explainable meaning - in contrast to q with respect to probabilities.

Second, in contrast to Dash et al,[12] the calculated entropy values were not normalized. The entropy measure depends on the number of distances calculated for a dataset. We thus expect higher entropy values in larger datasets. However, because entropy values are compared only within the same number of dimensions of a dataset, normalization of the calculated entropy values is not necessary here. Normalization of the entropy value might even be misleading, because the entropy then no longer indirectly reflects the number of samples of a dataset and the number of dimensions. However, different dimensions still might not be compared with each other.

Third, Dash et al [12] used two additional correction parameters, meeting point μ and coefficient β. Coefficient β was chosen to correct for a rapidly increasing entropy in the case of very small distances leading to quite different entropy values calculated for distances within a cluster. There was no clear decision rule on which parameter value should be chosen for β, but the parameter value was chosen with respect to the resulting entropy curve. Dash et al[12] suggested a value around 10, which seems to work well in their study.[12] The meeting point μ should help to differentiate between intra- and inter-cluster distances more accurately. It might be difficult to distinguish intra- and inter-cluster distances if the distance is 0.5. The parameter μ was calculated as 0.185 based on parameter β.[12] Both parameters can help to correct the entropy measure. However, both μ and β must be estimated or set to a (subjective) value. This might consequently introduce an additional inaccuracy. Thus, we did not include these two parameters in our formula (Equation (2)).

The described procedure was performed for all possible combinations of isotopic systems. The number of possible combinations can be calculated by

$$2^k - 1 \tag{5}$$

where k is the total number of isotopic systems.

All statistical and data mining analyses were performed using the R software.[28] The R programming code used for entropy calculation is available in the supporting information S2.

2.3 | Gaussian Mixture Model clustering

Gaussian Mixture Model (GMM) clustering was performed using the package "mclust" version 5.3 within the statistical program "R".[17,28]

To compare entropy-based feature ranking (see above) and clustering results multivariate outliers were removed from the data as described earlier. Furthermore, isotopic data were normalized according to Equation (3) prior to clustering as the different ranges within the isotope systems under study would even have a considerable influence on clustering when comparing combinations of isotopic systems. The normalized dataset was clustered for all possible combinations of isotopic systems ranging from one dimension to up to five dimensions depending on the subset.

The R package "clusterCrit" was used for calculating the trace index (see section 1.3).[29]

2.4 | Marginal and partial correlations

Correlations between isotopic systems may have an impact on feature ranking results. Two different types of correlation were conducted, namely marginal and partial correlations. The marginal correlation between two variables x_i and x_j is described by

$$r_{ij} = \frac{\mathrm{cov}\left(x_i, x_j\right)}{\sqrt{\mathrm{var}\left(x_i\right)}\sqrt{\mathrm{var}\left(x_j\right)}}. \tag{6}$$

The partial correlation of variables x_i and x_j while controlling for x_k can be calculated by the following equation:

$$r_{ij \cdot k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}} \tag{7}$$

Partial correlation describes the relationship between two (random) variables after removing the effect of all other (random) variables. Thus, partial correlation only gives the "unaffected" actual correlation between two variables, without the potential influence of another variable, which was removed from the dataset. Accordingly, partial correlations might be of interest if variables are removed as performed in this study.

Correlation analyses were conducted using R software.[28] The R package "ppcor" was used to calculate the partial correlation.[30]

3 | METHOD EVALUATION

Application of the feature ranking method described earlier strongly depends on the basic question of a research study that should be answered using stable isotopic ratios. In our study we illustrate three possible subsets with different issues and a varying number of isotopic systems to demonstrate the effect of isotopic ranking using entropy.

The entropy values differ between the datasets using four (dataset I) and five (dataset II) dimensions due to the different species included in the dataset as well as differences in sample numbers (see section 2.1).

Furthermore, we must point out that it is not possible to rank features across dimensions, as distance-based entropy mathematically increases with an increasing number of dimensions. Thus, higher entropy values detected, for example, in the two-dimensional subset than in the one-dimensional subset are mathematical artifacts since more distances are computed in the two-dimensional subset.

3.1 | Evaluation of the feature ranking method

The applied feature ranking method (without preliminary outlier removal) was tested using different artificially generated test sets (T1–T4) with certain properties as illustrated in Figure S1 and Table S1 (supporting information S1). Test set T1 contained one outlier value in variable B1 to test for the influence of extreme values. Test set T2 contained an outlier for both variables A2 and B2, while variable C2 was grouped into two quite distinct clusters. It is important to test if a variable including an outlier would be preferred over another variable, which is clustered into clear groups. Variables A3 and B3 of test set T3 both could be separated into three clusters, where B3 showed a more distinct separation of the clusters. Variable C3, however, did not show any clustering and a relatively high standard deviation. Finally, test set T4 gave a variable separated into two clusters (A4), another variable exhibiting three groups (B4), and a variable without any clustering structure (C4) (Figure S1 and Table S1, supporting information S1).

When the feature ranking method was applied on the four different test sets, the following results were obtained. For test set T1, where a single outlier was found in variable B1, ranking indeed resulted in the lowest entropy values for B1. This was clearly caused by the outlier, which is the only difference between variables A1 and B1 (Figure S2A, supporting information S1). Consequently, the exclusion of (multivariate) outliers, as described in the Experimental section, is recommended. Entropy did not find any difference between variables A1 and B1 after the removal of the outlier (not shown in this study). Test set T2, including the same single outlier in variables A2 and B2, however, with an additional grouping into two clusters in the case of variable C2, showed that an outlier (at least a single outlier value) does not have the ability to affect feature ranking in the presence of a clearly structured variable. Variable C2 exhibited the lowest entropy values (Figure S2B, supporting information S1). Nevertheless, outlier exclusion might be recommended. In the third test set (T3) with no clustering structure for variable C3 but three clusters in both A3 and B3 with a clearer structuring of the latter one; the lowest entropy value was found for the well-structured variable B3 as well as a combination of variables A3 and B3. Variable C3 showed a high entropy value because of its unstructured ("chaotic") distribution (Figure S2C, supporting information S1). Test set T4 was used to evaluate the feature ranking method when a different number of clusters was present in the dataset dependent on the variable. As expected, the separation into three clusters (B4) was favored over a separation into two groups (A4), consequently resulting in low entropy. Variable C4, lacking any clustering structure, again showed a high entropy measure. As for test set T3, the combination of variables A4 and B4 showed the lowest entropy values in the two-dimensional case (Figure S2D, supporting information S1).

For all tested combinations, feature ranking gave the expected result. As already mentioned, we clearly recommend removing (multivariate) outliers prior to feature ranking to avoid biased results.

3.2 | Composition of datasets-an evaluation

In the following sections different possible scenarios with varying underlying scientific questions were exemplarily tested. Different datasets were established with respect to isotopic systems used for feature ranking. Three different subsamples (terrestrial mammals, herbivorous mammals, fish) with four (dataset I: $\delta^{13}C_{collagen}$, $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$, $\delta^{18}O_{carbonate}$) and five (dataset II: $\delta^{13}C_{collagen}$, $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$, $\delta^{18}O_{carbonate}$, $\delta^{18}O_{phosphate}$) isotopic dimensions, respectively, were considered. For an easier labeling of feature combinations the isotopic dimensions were named as follows: $\delta^{13}C_{collagen} = 1$, $\delta^{15}N_{collagen} = 2$, $\delta^{13}C_{carbonate} = 3$, $\delta^{18}O_{carbonate} = 4$, and $\delta^{18}O_{phosphate} = 5$. For example, a subset including $\delta^{13}C_{collagen}$, $\delta^{15}N_{collagen}$, and $\delta^{13}C_{carbonate}$ values was called "123," while a subset including $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$, $\delta^{18}O_{carbonate}$, and $\delta^{18}O_{phosphate}$ values was named "2345."

We evaluated the entropy-based feature ranking method using a subset of our four-dimensional dataset I, including herbivorous, carnivorous, and omnivorous terrestrial mammals. As the relative frequency of herbivores, carnivores, and omnivores may vary between different datasets, we tested the influence of dietary groups on feature ranking results. Furthermore, a varying number of data in the subsets allowed for a validation of the feature ranking method dealing with smaller and larger sample sizes.

For each of the eight combinations of herbivores, carnivores, and omnivores tested in this study (evaluation sets A – H, Table S2, supporting information S1), ten sample sets of a given sample size (n = 40 – 80; see Table S2, supporting information S1) and of a given absolute ratio of herbivores, carnivores, and omnivores (see Table S2, supporting information S1) were randomly drawn from the whole dataset I. To avoid a bias, multivariate outliers were removed from each subset (herbivores, carnivores, and omnivores) separately. After the removal of outliers, the whole evaluation dataset (n = 92) consisted of a total of 55 herbivores, 21 carnivores, and 16 omnivores. The distance-based entropy was calculated for each feature combination as described earlier.

Several subsets, including different proportions of herbivores, carnivores, and omnivores, were tested as shown in Table S2 (supporting information S1). The feature ranking result shown in Tables S2 and S3 (supporting information S1) was the most frequent of the ten conducted runs conducted of the ten sample sets tested in each dimension. Feature ranking showed some variability as a consequence of the varying ratio of herbivorous, carnivorous, and omnivorous terrestrial mammals (Table S2, supporting information S1). Two different scenarios were detected: $\delta^{15}N_{collagen}$ values were best in separating the dataset in half of the tested sets, while $\delta^{13}C_{carbonate}$ values were able to optimally separate the dataset in the other half. The respective other isotopic system was then found on the second rank. In all tested scenarios $\delta^{18}O_{carbonate}$ values resulted in the highest entropy values in the one-dimensional case (Table S2, supporting information S1).

While $\delta^{15}N_{collagen}$ values showed the lowest entropy value if the proportion of herbivores was distinctly higher than that of both carnivores and omnivores, $\delta^{13}C_{carbonate}$ values became more important if the proportion of herbivores declined and was thus more similar to the relative frequency of both carnivores and omnivores (Table S2, supporting information S1).

In the two-dimensional case (Table S3, supporting information S1), the combination of $\delta^{13}C_{carbonate}$ values with both $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ values resulted in the lowest entropy values. In three sets (D, F, and H) an equal number of random samples showed lowest entropy values for "13" and "23." Combining $\delta^{13}C_{collagen}$ and $\delta^{18}O_{carbonate}$ values resulted in the highest entropy values, and thus the lowest information content, in all but one (E) set. This seems to be in good accordance with the worst entropy results using one isotopic system only (Table S3, supporting information S1).

Removing a single isotopic system, consequently leading to a three-dimensional dataset, resulted in three different optimal combinations, depending on the proportion of herbivores, carnivores, and omnivores, namely "123," "134," and "234." Interestingly, the combination of $\delta^{13}C_{collagen}$, $\delta^{13}C_{carbonate}$, and $\delta^{18}O_{carbonate}$ values ("134") resulted in the lowest, thus best, entropy values in sets F and G, while entropy was maximal, thus worst, for sets B and C. For all other sets, removing $\delta^{13}C_{carbonate}$ values from the data resulted in the highest entropy values (Table S3, supporting information S1) and is thus not recommended.

Consequently, dependent on the composition of the dataset under study we expect varying feature ranking results.

4 | RESULTS

4.1 | Correlation between the isotopic systems

The results revealed by correlation analysis are shown in Tables S5 and S6 (supporting information S1) for datasets I and II, respectively. The different subsets chosen for feature ranking (see below) showed several significant correlations between the isotopic systems for both marginal and partial correlations.

No overall pattern could be detected with respect to correlation. In the present study only the $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ values showed significant marginal and partial correlations for almost all subsets in both datasets tested. Only the subset of herbivorous mammals showed some variability here with a nonsignificant marginal correlation as well as a negative marginal and partial correlation coefficient in both datasets I and II (Tables S5 and S6, supporting information S1).

Terrestrial and herbivorous mammals showed a significant correlation, both marginal and partial, between $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ values. Although these two isotopic systems were derived from the same material (bone carbonate), no significant relationship could be detected in the fish subset (Tables S5 and S6, supporting information S1).

In dataset II (including $\delta^{18}O_{phosphate}$) both terrestrial and herbivorous mammals showed a significant (marginal and partial) correlation between $\delta^{18}O_{carbonate}$ and $\delta^{18}O_{phosphate}$ values (Tables S5 and S6, supporting information S1).

The correlations detected between the isotopic systems might play an important role for feature ranking (see below).

4.2 | Terrestrial mammals

If one considers a typical archaeological site, terrestrial mammals would probably represent the majority of animal bone finds including both wild and domesticated individuals. Scientific questions could include the detection of non-local (imported) individuals, differences in food supply, and water sources.

The subsets containing 99 (dataset I) and 91 (dataset II) terrestrial mammals can be understood as exemplary datasets for terrestrial wild and domesticated mammals of a variety of different species. The relative frequencies of herbivores, carnivores, and omnivores for both subsets are shown in Table S7 (supporting information S1). The proportion of herbivorous, carnivorous, and omnivorous mammals was quite similar in the analyzed datasets I and II (Table S7, supporting information S1); therefore, we might expect similar feature ranking results. In both datasets, the proportion of herbivores was markedly higher than those of both carnivores and omnivores (Table S7, supporting information S1). Furthermore, the proportion of herbivores, carnivores, and omnivores was quite similar to that in our evaluation set C (see section 3.2; Table S2, supporting information S1). In addition, we expect a separation of the isotopic data with respect to diet because of the inclusion of herbivores, carnivores, and omnivores in this subset.

The feature ranking results of the terrestrial mammals are illustrated in Figure 1 and Table 1. Using only one isotopic system, $\delta^{15}N_{collagen}$ values were preferred, resulting in the lowest entropy values. The $\delta^{13}C_{carbonate}$ values also showed low entropy values in both datasets. In addition, combinations of other isotope ratios with $\delta^{15}N_{collagen}$ or $\delta^{13}C_{carbonate}$ values seem to result in relatively low entropy values compared with other combinations, especially combinations including $\delta^{18}O_{carbonate}$ values. Moreover, combining both $\delta^{15}N_{collagen}$ and $\delta^{18}O_{carbonate}$ values resulted in more intermediate entropy values. However, combinations of other isotopes with $\delta^{13}C_{collagen}$ and $\delta^{18}O_{phosphate}$ values showed rather high entropy values. As expected, both datasets gave corresponding results for one to four dimensions with respect to minimum entropy values (see Table 1). In the case of dataset the II $\delta^{18}O_{phosphate}$ values and combinations of the latter with other isotopic systems resulted in (slightly) lower entropy values than with $\delta^{18}O_{carbonate}$ values (Figure 1 and Table 1). Thus, when only the two oxygen isotopic systems were compared, $\delta^{18}O_{phosphate}$ values would be preferred. Consequently, with regard to the subset of terrestrial mammals the ranking of both datasets I and II indicated that the removal of $\delta^{18}O_{carbonate}$ values would cause no or only minor loss of information.

The feature ranking results should also be visible in the clustering outcome. Clustering the normalized data using GMM showed the following results: In the four-dimensional dataset, clustering all

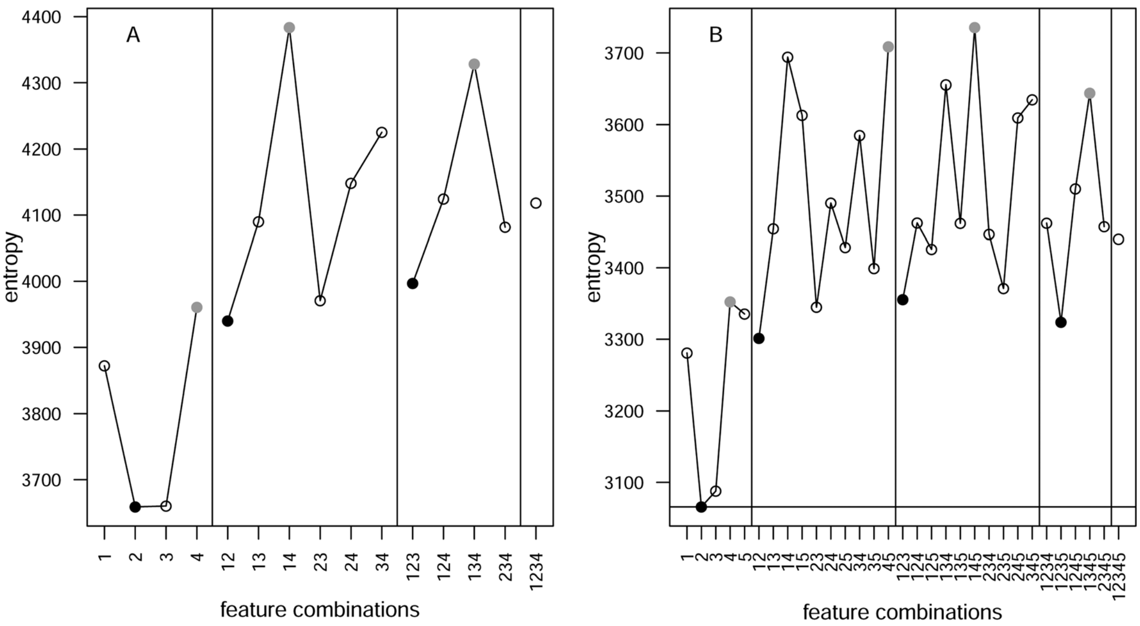

FIGURE 1 Calculated entropy values for the possible combinations of isotopic systems in A, dataset I and B, dataset II for terrestrial mammals from Haithabu and Schleswig. In each dimension (separated by vertical lines) the minimal (filled black point) and maximal (filled gray point) entropy values are highlighted. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$, 5 = $\delta^{18}O_{phosphate}$

TABLE 1 Feature ranking results (minima and maxima) as well as the optimal trace index (maxima) for the terrestrial mammals from Haithabu and Schleswig in datasets I and II for each dimension. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$, 5 = $\delta^{18}O_{phosphate}$

Feature ranking using entropy				
Minima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	2	12	123	
II	2	12	123	1235
Maxima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	4	14	134	
II	4	45	145	1345
Trace index				
Maxima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	2	12	123	
II	2	12	123	1235

dimensions only slightly differed from clustering without $\delta^{18}O_{carbonate}$ values ("123") (Figures S3 and S4, supporting information S1). Both clustering processes resulted in four clusters, with only eight individuals grouped into different clusters when comparing the clustering results, namely cattle Hb 54, Hb 56, Hb 59, and Hb 60, horse Hb 91, roe deer S 10, and sheep Hb 74 and Hb 76 (Table S9, supporting information S1).

Regarding dataset II, clustering without $\delta^{18}O_{carbonate}$ values ("1235") resulted in three instead of four clusters in the five-isotope ("12345") scenario (Figures S5 and S6, Table S10, supporting information S1). Consequently, in this case at least some information loss can be observed.

These findings were in accordance with the cluster validation results using the trace index with an optimal trace index for exactly those combinations of isotopic systems resulting in optimal (minimum) entropy values (see Table 1). Therefore, we expect that combinations of isotopic dimensions exhibiting the lowest entropy values also result in good clusters due to their optimal trace index as demonstrated by GMM clustering.

4.3 | Herbivorous mammals

Multi-isotope analyses of a dataset consisting of only herbivorous mammals could, for example, help to detect primarily non-local individuals at the study site or give hints at different food and water sources of, for example, wild and domesticated mammals.

Feature ranking on herbivorous mammals (dataset I: *n* = 55, dataset II: *n* = 49) led to a shift compared with the ranking of all terrestrial mammals, even including carnivores and omnivores (see above). $\delta^{15}N_{collagen}$ values became less important for the herbivorous dataset than for the dataset including herbivores, carnivores, and omnivores (Figure 1 and Table 1). On the contrary, $\delta^{13}C_{carbonate}$ values as well as combinations of $\delta^{13}C_{carbonate}$ values with the other isotopic systems became more important, resulting in relatively low entropy values. However, when only one isotopic system must be removed, the exclusion of $\delta^{18}O_{carbonate}$ values is recommended in both datasets (Figure 2 and Table 2), as was the case for all terrestrial mammals (see section 4.2).

When clustering the herbivores of dataset I without $\delta^{18}O_{carbonate}$ values ("123"), only two instead of three clusters as in the four-dimensional dataset can be detected (Figures S9 and S10, supporting information S1). Two clusters of the whole dataset were grouped together (Table S10, supporting information S1). Again, the removal of $\delta^{18}O_{carbonate}$ values is connected to a loss of information.

GMM clustering resulted in three clusters when clustering all five dimensions as well as clustering without $\delta^{18}O_{carbonate}$ values ("1235"). Besides the fact that the clusters were (automatically) numbered differently in these two scenarios ("12345" vs. "1235": cluster 1 = cluster 3, cluster 2 = cluster 1, cluster 3 = cluster 2), the detected clusters were quite similar in both cases. However, ten individuals (aurochs Hb 48, cattle Hb 57, hare Hb 2 and S 24, horse Hb 86, Hb 89, and Hb 90, red deer Hb 35 and S 1, and sheep Hb 79) were grouped into a different cluster when comparing the two scenarios (Table S11, supporting information S1).

Similar to feature ranking, the cluster validation using the trace index showed best results when using $\delta^{13}C_{carbonate}$ values or

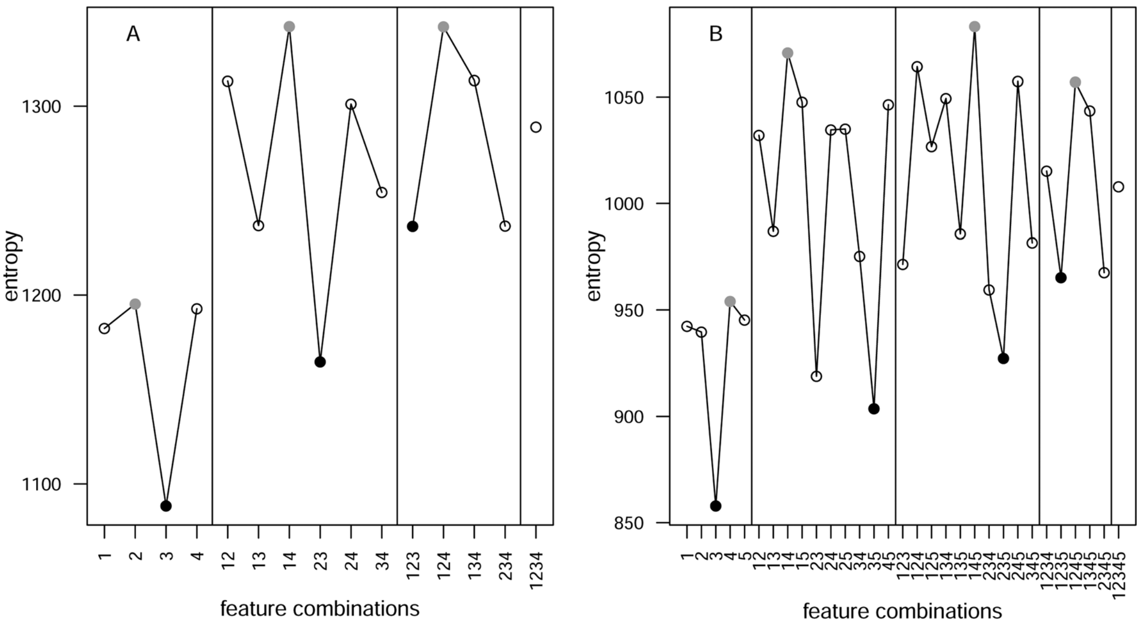

FIGURE 2 Calculated entropy values for the possible combinations of isotopic systems in A, dataset I and B, dataset II for terrestrial herbivorous mammals from Haithabu and Schleswig. In each dimension (separated by vertical lines) the minimal (filled black point) and maximal (filled gray point) entropy values are highlighted. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$, 5 = $\delta^{18}O_{phosphate}$

TABLE 2 Feature ranking results (minima and maxima) as well as the optimal trace index (maxima) for the terrestrial herbivorous mammals from Haithabu and Schleswig in dataset I and II for each dimension. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$, 5 = $\delta^{18}O_{phosphate}$

Feature ranking using entropy				
Minima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	3	23	123	
II	3	36	235	1235
Maxima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	2	14	124	
II	4	14	145	1245
Trace index				
Maxima	**1 dimension**	**2 dimension**	**3 dimension**	**4 dimension**
I	3	34	234	
II	3	13	235	1235

combinations including $\delta^{13}C_{carbonate}$ values. Furthermore, in the case of dataset II both feature ranking and trace index chose “235” and “1235” as the best combinations of three and four isotopic systems, respectively (Table 2).

4.4 | Fish

The fish subset (*n* = 46) showed low entropy values when analyzing $\delta^{13}C_{collagen}$ values only as well as combinations of isotopic systems including collagen carbon isotope ratios (Figure 3 and Table 3). As for the herbivorous subset, combinations with $\delta^{15}N_{collagen}$ values resulted in rather high entropy values, probably due to the poor information content of this isotopic system. Using the information of more dimensions, especially the combination of $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ values exhibited a low entropy value (Figure 3 and Table 3). According to the feature ranking results, the removal of $\delta^{15}N_{collagen}$ values should cause a rather small loss of information (Figure 3 and Table 3).

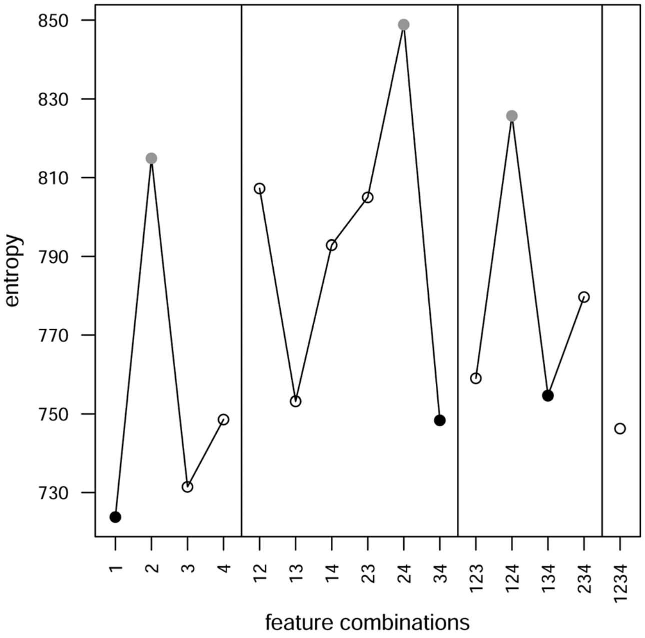

FIGURE 3 Calculated entropy values for the possible combinations of isotopic systems in dataset I for fish from Haithabu and Schleswig. In each dimension (separated by vertical lines) the minimal (filled black point) and maximal (filled gray point) entropy values are highlighted. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$

GMM clustering of all isotopic dimensions compared with the clustering without $\delta^{15}N_{collagen}$ values (“134”) showed two identical clustering results with two relatively distinct clusters with the exception of only two individuals (perch 48FB5Pop, pike 10H1C; Table S12, supporting information S1).

A previous cluster analysis (without data normalization) of the fish dataset from Haithabu and Schleswig revealed an optimal number of four clusters, namely a freshwater cluster (cluster 3), a brackish water cluster (cluster 4), and two marine clusters (clusters 1 and 2; see Table S13, supporting information S1).[4] Freshwater, brackish, and marine clusters were mainly separated from each other due to their $\delta^{13}C_{carbonate}$ values. However, individuals of cluster 1 were enriched in ^{18}O compared with all other clusters, indicating an origin from another, colder environment. Thus, fish from the first cluster were probably non-local to our study sites Haithabu and Schleswig.[4] Table S13 (supporting information S1) shows the comparison of the cluster results using the whole fish dataset without normalization according to Göhring et al.[4] and the results of the clustering of the carbonate fraction after normalization (this study). Interestingly, clustering with only two dimensions, where entropy was optimal when selecting $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ values only (“34”, Figure 3 and Table 3), was very similar to the clustering of the whole dataset without prior normalization of the data.[4] This was especially conspicuous for the cluster including probably non-local fish (cluster 1 in case of “34”), which was the most important cluster when the task was to detect primarily non-local individuals (Table S13, supporting information S1; see Göhring et al.[4]). Cluster 1 differs from the previous cluster 1 by only three individuals. Two individuals (cod 1D1V and cod 4D4V) were previously grouped into the marine cluster 2, and another cod (cod 42D4V) was previously grouped into cluster 1. However, when clustered using the (normalized) carbonate fraction only (“34”), this individual was grouped into the marine cluster (cluster 3 in case of “34”; Table S13, supporting information). Cluster 2 combined the previous freshwater cluster (cluster 3 in the case of the not-normalized “1234”) and parts of the brackish water cluster 4. The remaining brackish water individuals were grouped into the fourth cluster in the case of “34” (Table S13, supporting information

TABLE 3 Feature ranking results (minima and maxima) as well as the optimal trace index (maxima) for the fish from Haithabu and Schleswig in dataset I for each dimension. 1 = $\delta^{13}C_{collagen}$, 2 = $\delta^{15}N_{collagen}$, 3 = $\delta^{13}C_{carbonate}$, 4 = $\delta^{18}O_{carbonate}$

Feature ranking using entropy			
Minima	**1 dimension**	**2 dimension**	**3 dimension.**
I	1	34	134
Maxima	**1 dimension**	**2 dimension**	**3 dimension.**
I	2	24	124
Trace index			
Maxima	**1 dimension**	**2 dimension**	**3 dimension.**
I	1	13	123

S1). The differences between the two cluster analyses are rather small, probably as an effect of data normalization. As already mentioned, clustering with the carbonate fraction only (“34”) revealed the cluster of probably non-local fish, which can be considered one of the main goals of Göhring et al.[4]

The trace index coincided with the ranking results only in the one-dimensional case. For both two- and three-dimensional feature combinations, the trace index was equal to the second-best feature ranking result (“13” and “123”; Figure 3 and Table 3).

5 | DISCUSSION

5.1 | Evaluation of the feature ranking method

The evaluation procedure using isotopic data demonstrated that the ranking procedure is, logically, dependent on the composition of the dataset. While $\delta^{15}N_{collagen}$ values were most important for structuring the dataset when the proportion of herbivores was higher than that of carnivores or omnivores, $\delta^{13}C_{carbonate}$ values became more relevant when the proportion of herbivores (as well as the absolute number of herbivores) decreased. Nitrogen stable isotope ratios reflect the dietary protein. Consequently, it comes as no surprise that herbivores can be best separated from both carnivores and omnivores according to their $\delta^{15}N_{collagen}$ values. However, with a more similar ratio of herbivores, carnivores, and omnivores, $\delta^{15}N_{collagen}$ values become less relevant than $\delta^{13}C_{carbonate}$ values. The latter values are capable of separating herbivores, carnivores, and omnivores according to the overall composition of their diet. It is important to mention that the total number of individuals in those test sets resulting in the lowest entropy values for $\delta^{13}C_{carbonate}$ (sets D, F, G, and H) was relatively low, with only 40 and 48 individuals (Table S2, supporting information S1). Thus, the shift in the ranking could also be caused by the rather low sample sizes. However, test set E had a similar (but slightly higher) sample size ($n = 52$), and the feature ranking method still showed the lowest entropy values for $\delta^{15}N_{collagen}$ (Table S2, supporting information S1). Nevertheless, it is advisable to perform the feature ranking method with a higher number of data to gain trustworthy results.

Entropy-based feature ranking was compared with the trace index, which is a cluster validation index. Clustering should be optimal when the trace index is maximum. Moreover, an optimal clustering, in the sense of clearly structured data points, should result in the lowest entropy values. Thus, we expect similar results for ranking and validation.

The optimal feature combinations according to the trace index were identical for the subset of terrestrial mammals (Table 1). Some variations were present in the subsets of herbivores and fish (see Tables 2 and 3). These differences can be explained as follows: The entropy values of feature combinations “123” and “234” in dataset I of the subset of terrestrial herbivores were almost identical with slightly better results for “123” (see Figure 2A). However, the removal of $\delta^{18}O_{carbonate}$ values (“123”) did not result in the optimal clustering structure according to the trace index. Similarly, as regards the fish subset feature combinations “34” and “134” were classified as optimal with respect to their entropy values. However, even the combinations of “13” and “123” resulted in relatively low entropy values, thus indicating a quite well-structured dataset, namely the second-best ranking results for the two- and three-dimensional sets, respectively. Indeed, both “13” and “123” were optimal according to the trace index. In addition, herbivores showed differences in the two-dimensional ranking of both datasets I and II. While both these combinations were rather different from the optimal combination with respect to entropy, they showed relatively similar results for the trace index (not shown in this study). This could explain the divergences between entropy-based feature ranking and trace index.

5.2 | Terrestrial mammals

For the subset including all terrestrial mammals, entropy-based feature ranking pointed towards relatively high information content, especially as regards both $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ values. Since the carbon and nitrogen isotope ratios of bone collagen are related to the protein part of the diet, it comes as no surprise that a dataset including a mixture of herbivores, carnivores, and omnivores (Table S7, supporting information S1) can be best separated according to these isotopic systems.

On the contrary, the information provided by $\delta^{18}O_{carbonate}$ values was not sufficient, resulting in quite high entropy values (Figure 1 and Table 1). In addition, the $\delta^{18}O_{phosphate}$ values were evaluated as more important than the $\delta^{18}O_{carbonate}$ values in the five-dimensional dataset. Consequently, the exclusion of $\delta^{18}O_{carbonate}$ values would not result in a loss of much information. However, we must strongly emphasize that this can definitely not be interpreted in the sense that $\delta^{18}O_{carbonate}$ values do not contain any information at all. Indeed, it only contained less information than other isotopic systems in our datasets. In addition, some information loss was observed in dataset II, resulting in three instead of four clusters when removing $\delta^{18}O_{carbonate}$ values (“1235”; Figures S5 and S6, Table S9, supporting information S1).

The relatively low information content detected for $\delta^{18}O_{carbonate}$ values can probably be explained by a quite strong “sea spray” effect in this isotopic system as detected by Göhring et al.[31] This effect causes a distinct enrichment in ^{18}O in terrestrial mammals from Haithabu and Schleswig, thus leading to an overlap of herbivores, carnivores, and omnivores. Although this effect has also been verified for $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{phosphate}$ values, it is much stronger in $\delta^{18}O_{carbonate}$ values.[31,32] Consequently, the high entropy values for $\delta^{18}O_{carbonate}$ might be a site-specific result and must be verified for other datasets.

5.3 | Herbivorous mammals

Terrestrial herbivores were best separated by $\delta^{13}C_{carbonate}$ values and their combinations with other isotopic systems. The information contained in $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ values (see above) was no longer very important in a subset consisting of herbivores only. $\delta^{18}O_{carbonate}$ values also seem to play a minor role for the herbivorous subset resulting in relatively high entropy values (Figure 2 and Table 2). However, even in this case this is not equivalent to a meaningless isotopic system. Moreover, GMM clustering showed a loss of information when $\delta^{18}O_{carbonate}$ values were removed from dataset I ("123") compared with the complete dataset (Figures S7 and S8, Table S10, supporting information). This, consequently, confirms that the suggested feature removal with respect to the entropy measure is clearly not equivalent to the detection of a meaningless isotopic system. It must rather be understood as the feature that would cause the lowest loss in information when being removed from the dataset.

It is important to mention that feature combination "123" (dataset I) and "1235" (dataset II) exhibit only slightly lower entropy values than the "234" (dataset I) and "2345" (dataset II), respectively. Thus, the removal of $\delta^{18}O_{carbonate}$ values and that of $\delta^{13}C_{collagen}$ values cause almost the same loss of information. Entropy-based feature ranking prefers to remove $\delta^{18}O_{carbonate}$ values, while the trace index prefers to remove $\delta^{13}C_{collagen}$ values in the case of dataset I. Indeed, GMM clustering without $\delta^{13}C_{collagen}$ values ("234") results in three clusters (not shown in this study) instead of only two after the removal of $\delta^{18}O_{carbonate}$ values ("123"). For dataset II, both feature ranking and trace index are in agreement in removing $\delta^{18}O_{carbonate}$ values (Figure 2 and Table 2). Consequently, here the decision on the removal of an isotopic system also relies on the underlying scientific question to be solved.

5.4 | Fish

Entropy-based feature ranking showed that $\delta^{13}C_{collagen}$ values, as well as combinations of especially $\delta^{18}O_{carbonate}$ values and $\delta^{13}C_{carbonate}$ values, are relatively important in the fish subset (Figure 3 and Table 3). Differentiation between marine and freshwater fish is, among others, possible using $\delta^{13}C_{collagen}$ values.[33] However, the information content contained in $\delta^{15}N_{collagen}$ values was comparatively low.

Comparison of the GMM clustering of the total fish data ("1234") and the dataset without $\delta^{15}N_{collagen}$ ("134") showed almost no differences, with the exception of the clustering results of two single individuals (Figures S11 and S12, Table S12, supporting information). Therefore, the exclusion of $\delta^{15}N_{collagen}$ data would result in almost no information loss.

Since $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ data are usually generated in parallel, even the exclusion of both dimensions must be investigated. This might be necessary if collagen is not (well) preserved in fish remains. Indeed, clustering with the carbonate fraction only ("34") led to the detection of four (compared with two) distinct clusters (Figure S13, Table S13, supporting information). Interestingly, clustering of the four-dimensional (not-normalized) fish data showed similar clustering results, among others, indicating a cluster of non-local individuals.[4] Therefore, the information necessary for the detection of a group of probably non-local individuals was almost solely present in this two-dimensional dataset consisting of $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ values only. Consequently, the collagen fraction did not play a major role for the fish data when the main aim was to detect primarily non-local individuals. This comes ′as no surprise as the stable carbon and oxygen isotope ratios of the carbonate fraction both indicate a salinity and a temperature signa[3,34-37] which caused the main cluster structure in the original study (see above and Göhring et al.[4]) Furthermore, non-local individuals, which originated from a colder environment in the case of our dataset, could be identified due to their $\delta^{18}O_{carbonate}$ values.[4]

5.5 | Correlation between isotopic systems

Depending on the subset different isotopic systems resulted in relatively high entropy values and, consequently, they could be removed from a dataset without too much loss of information. It is indeed possible to detect a relationship between the less informative isotopic system in each subset and the correlation between the isotopic systems (Tables S5 and S6, supporting information). For all subsets examined in this study, the isotopic system, which was proposed to be omitted according to entropy-based feature ranking, was linked to at least two other isotopic systems by a significant correlation. At least one of these relationships also showed a significant partial correlation (Tables S5 and S6, supporting information).

The subset of terrestrial mammals showed a tendency to omit $\delta^{18}O_{carbonate}$ values in datasets I and II. $\delta^{18}O_{carbonate}$ values were significantly correlated to both $\delta^{13}C_{collagen}$ and $\delta^{13}C_{carbonate}$ values in dataset I and to $\delta^{13}C_{collagen}$, $\delta^{13}C_{carbonate}$, and $\delta^{18}O_{phosphate}$ values in dataset II.

Terrestrial herbivores showed the same correlation results in both datasets. However, regarding the herbivorous subset, the correlation between $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ values was conspicuous due to a negative correlation coefficient. Lower $\delta^{13}C_{collagen}$ values were combined with higher $\delta^{15}N_{collagen}$ values (and the other way around). This might have been caused by a limnic influence detected for the nitrogen stable isotopic system causing an enrichment in ^{15}N in some herbivores in our dataset[38] while $\delta^{13}C_{collagen}$ values were not affected. This effect was certainly also found in the terrestrial subset, however, probably masked by the presence of carnivores and omnivores.

The removal of nitrogen isotope ratios did obviously not cause a loss in information in the fish subset. $\delta^{15}N_{collagen}$ values correlated with $\delta^{13}C_{collagen}$, $\delta^{13}C_{carbonate}$, and $\delta^{18}O_{carbonate}$ values. In addition, both $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ values were significantly correlated with $\delta^{13}C_{collagen}$ and $\delta^{15}C_{collagen}$ values with respect to the marginal correlation and with $\delta^{13}C_{collagen}$ values in the partial correlation. Thus, a sufficient amount of information contained in the collagen fraction was also contained in the carbonate fraction.

Consequently, variables that show high marginal and especially partial correlation factors with other variables in the dataset can be more easily omitted from the dataset without losing too much information. The remaining correlated variables are obviously capable of at least partly replacing the information content of the removed variable.

5.6 | Applicability of the entropy-based feature ranking method

In order to reduce costs or sample material needed for isotope analysis, isotopic systems extracted and analyzed together (e.g., $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$, $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$) must also be excluded from the analysis together. This is also considered in the fish subset (see above). Our method even allows the entropy of the different feature combinations after the removal of two or more isotopic systems to be considered. Consequently, if the material is poorly preserved or available in a limited amount, the task is to decide which isotopic systems must be analyzed and which ones should be excluded from analyses. $\delta^{13}C_{collagen}$ and $\delta^{15}N_{collagen}$ or $\delta^{13}C_{carbonate}$ and $\delta^{18}O_{carbonate}$ then must be removed pairwise in order to reduce the necessary amount of sample material. Nevertheless, even the information content of other stable isotopic systems could be evaluated using our method, for example, $\delta^{34}S_{collagen}$ and $^{87}Sr/^{86}Sr$. In principle, our feature ranking method can also be used for a comparison of one-dimensional features, i.e. single isotopic systems. However, the information content of a single variable is usually limited anyway. Analysis of more isotopic systems would, however, necessitate additional sample material, which might be limited. Consequently, it must be considered if an additional isotopic system would result in information gain. Depending on the research question, feature ranking based on entropy would allow validation of the information loss accompanied with the removal of certain isotopic systems. Moreover, the method also allows the feature combination, which results in the second lowest (second best) entropy value to be chosen, if the optimal feature combination is not possible due to, for example, insufficiently preserved sample material. We would, however, like to emphasize that a multi-isotope approach is (usually) clearly recommended wherever possible. Recent studies demonstrated that the analysis and interpretation of multiple isotopic systems result in a gain of information compared with common bivariate analyse.[4-7,31,32] Consequently, the exclusion of certain isotopic systems from analysis will result in a certain loss of information. Thus, if sample material is available in sufficient quantity, one should choose the multi-isotope approach. However, as already mentioned, in the case of, for example, archaeological remains, sample material is often limited. Therefore, it might be necessary to reduce the number of analyzed isotopic systems and to select certain isotopic systems without losing (too much) information. Feature ranking based on entropy aims to reduce this loss of information by detecting the isotopic system(s) with the highest or lowest information content.

The application of our entropy-based feature ranking method in another study, which plans to use a multi-isotope approach on restricted sample material, requires the *a priori* decision which isotopic system(s) include(s) less information than others. For this purpose, a small but representative subset of the sample should be chosen as a pre-test. In this subset all different isotopic systems that are of interest for answering the research question(s) must be analyzed. Feature ranking should then be applied on the subset and accordingly allowing it to decided which isotopic system(s) could be omitted from the analyses of the remaining sample material without losing much information. Moreover, the results of other studies on similar sample material (e.g., fish bones) and maybe even similar hypotheses can also be used as a pretest for feature ranking. A higher similarity of, for example, the selected species will lead to more valid results. However, if no similar dataset is available, entropy-based feature ranking on the material under study using a small pre-test subset is the method of choice.

According to our results both the GMM cluster analysis and the trace index can be seen as a useful tool for the validation of the entropy-based feature ranking method. This might, however, usually not be a necessary step when applying the feature ranking method, but serves as a validation tool for implementing the entropy-based feature ranking described in this paper. Nevertheless, GMM cluster analysis was validated as a useful tool for the interpretation of multi-isotope datasets, previously.[4-6,8,31,32]

Finally, it is important to mention that our feature ranking method cannot be applied to studies where the underlying research question does not assume different groups (clusters) in the dataset. This is because the entropy would be lowest for a well-structured dataset with at least two clusters, but high for an unstructured dataset lacking any clusters. However, as mentioned before, many research questions related to stable isotope analyses aim to detect primarily non-local individuals, individuals with different diet or status, and individuals inhabiting different habitats or ecosystems. This would result in datasets with two or more clusters. Thus, entropy-based feature ranking is a valuable tool to validate the information content of different isotopic systems and to choose the feature combination with the highest information content if one or more isotopic systems must be excluded from analysis.

6 | CONCLUSIONS

Whenever possible, a multi-isotope approach should be preferred. It has been shown previously (see the Introduction section) that multi-isotope data analyses are part of future isotope studies. New data mining methods are therefore needed to analyze isotopic datasets.

However, especially in the case of archaeological studies, the material available for stable isotope analyses (e.g., bones, teeth, and hair) is often limited or certain skeletal components are insufficiently preserved. Therefore, it is of particular importance to decide which isotopic system(s) could be omitted without losing too much information. Entropy-based feature ranking offers a feasible and objective method to rank isotopes as well as a combination of isotopes and to select the isotopic systems that are most important to answer an underlying research question. Those isotopic systems that are less important according to entropy-based feature ranking of a pretest subset could be excluded from analyses to reduce the required material. In addition, this method can also be applied to modern specimen, reducing the amount of sample material (e.g., blood) in the case of live animals.

Our study showed that it is not possible to generally rank different isotopic dimensions without concerning the dataset. Feature ranking obviously depends on the composition of the dataset with respect to, for example, species- and diet-specific peculiarities. Terrestrial mammals, for example, showed a different ranking from herbivorous mammals only, or even fish. While $\delta^{18}O_{carbonate}$ values showed low information content in the subsets of both terrestrial mammals and herbivorous mammals, the isotopic system exhibiting the lowest information content with respect to entropy was collagen carbon for fish, where $\delta^{15}N_{collagen}$ values could have been excluded, although this in turn was the most important isotopic system for terrestrial mammals. Consequently, a general exclusion of a certain isotopic system could be highly erroneous. Nevertheless, multi-isotope analyses on a small (representative) pre-test subset will allow ranking of the isotopic systems of the whole material under study. The material needed for the analyses of the remaining majority of the sample can consequently be reduced based on the feature ranking results of the pre-test. The present study can be used as a first hint when investigating different groups of animals. In addition, even the multi-isotope data of other studies can be chosen for a pilot ranking as far as the investigated sample material is similar with respect to, for example, species as well as the research hypotheses.

In addition, we detected a relationship between the outcome of the entropy-based feature ranking and the correlation, both marginal and partial, between isotopic systems. Obviously, an isotopic system can be removed from a dataset without too much loss of information if the respective isotopic system was correlated with other isotopes in a sufficient amount (here: at least two marginal and one partial correlations). The information contained in the removed system is, accordingly, still at least partly present in the remaining isotopic systems and the removal does not cause a loss of information (or at least only a minor loss). Consequently, this might also be a first indication when deciding about the removal of an isotopic system.

We recommend using the described feature ranking method where no or only few data are available. However, a small (representative) subset of the collected material of a site should be analyzed as a pre-test and the tested isotopic systems should be ranked according to our method. These ranking results can be adopted on the remaining majority of the collected material. Site-specific differences in stable isotopes are probably also present in the feature ranking results. Further knowledge on feature ranking results from other sites is needed to detect potential general patterns in the isotopic data. This would, consequently, facilitate researchers to *a priori* decide on the set of isotopic systems that should be analyzed to gain as much information as possible when the available study material is limited. In addition, the removal of an isotopic system should certainly also be in accordance with the research question. Furthermore, it is important to emphasize that the described method also allows detection of the second-best combination of isotopic systems with respect to the entropy measure if, for example, gelatine could not be extracted. Consequently, entropy-based feature ranking can help to qualify even clustering of isotope ratios when one or more isotopic systems are not available for analysis.

ACKNOWLEDGEMENTS

Isotope analysis of bone collagen and carbonate was financially supported by the German Research Foundation DFG (Grant No. Gr959/15-1,2). Oxygen isotope analysis of bone phosphate was financially supported by the ArchaeoBioCenter of the Ludwig Maximilian University, Munich.

ORCID

Andrea Göhring https://orcid.org/0000-0002-5117-1948
Gisela Grupe https://orcid.org/0000-0002-2731-2296

REFERENCES

1. Kirsanow K, Tuross N. Oxygen and hydrogen isotopes in rodent tissues: Impact of diet, water and ontogeny. *Palaeogeogr Palaeoclimatol Palaeoecol*. 2011;310(1–2):9-16. https://doi.org/10.1016/j.palaeo.2011.03.022
2. Koch PL, Fogel ML, Tuross N. Tracing the diets of fossil animals using stable isotopes. In: Lajtha K, Michener RH, eds. *Stable Isotopes in Ecology and Environmental Science*. Oxford: Blackwell Scientific Publications; 1994:63-92.
3. Marshall JD, Brooks JR, Lajtha K. Sources of variation in the stable isotopic composition of plants. In: Michener R, Lajtha K, eds. *Stable Isotopes in Ecology and Environmental Science*. Malden, Oxford, Victoria: Blackwell Publishing; 2007:22-60.
4. Göhring A, Mauder M, Kröger P, Grupe G. Using Gaussian mixture model clustering for multi-isotope analysis of archaeological fish bones for palaeobiodiversity studies. *Rapid Commun Mass Spectrom*. 2016;30(11):1349-1360. https://doi.org/10.1002/rcm.7573
5. Grupe G, Klaut D, Mauder M, et al. Multi-isotope provenancing of archaeological skeletons including cremations in a reference area of the European Alps. *Rapid Commun Mass Spectrom*. 2018;32(19):1711-1727. https://doi.org/10.1002/rcm.8218
6. Mauder M, Ntoutsi E, Kröger P, Kriegel H.P. Towards predicting places of origin from isotopic fingerprints: A case study on the mobility of people in the central European Alps, In: Grupe G, McGlynn GC, eds. *Isotopic Landscapes in Bioarchaeology*. Berlin, Heidelberg: Springer; 2016:221–233. https://doi.org/10.1007/978-3-662-48339-8_13
7. Mauder M, Ntoutsi E, Kröger P, Kriegel H.P. The isotopic fingerprint: New methods of data mining and similarity search, In: Grupe G, Grigat A, McGlynn GC, eds. *Across the Alps in Prehistory: Isotopic Mapping of the Brenner Passage by Bioarchaeology*. Cham: Springer; 2017:105–126. https://doi.org/10.1007/978-3-319-41550-5_5
8. Mauder M, Ntoutsi E, Kröger P, et al. Applying data mining methods for the analysis of stable isotope data in bioarchaeology, *12th IEEE International Conference on e-Science, e-Science 2016*. Baltimore: IEEE; 2016:233–242. https://doi.org/10.1109/eScience.2016.7870904
9. Mauder M, Ntoutsi E, Kröger P, et al. Significance and limitations of stable oxygen isotope ratios in the apatite phosphate of archaeological vertebrate finds for provenance analysis in an Alpine reference region. *Archaeometry*. 2019;61(1):194-210. https://doi.org/10.1111/arcm.12399
10. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27(3):379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

11. Vajapeyam S. Understanding Shannon's Entropy metric for Information. *Comp Res Rep*. 2014;arXiv:1405.2061.
12. Dash M, Choi K, Scheuermann P, et al. Feature selection for clustering - a filter solution. In: *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society; 2002:115–122.
13. Bhowmik M, Sarkar A, Das R. Shannon entropy based fuzzy distance norm for pixel classification in remote sensing imagery. In: *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*. 2015. 1–6.
14. De Luca A, Termini S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Inf Control*. 1972;20(4):301-312. https://doi.org/10.1016/S0019-9958(72)90199-4
15. Liu R, Yang N, Ding X, Ma L. An unsupervised feature selection algorithm: Laplacian score combined with distance-based entropy measure, Third International Symposium on Intelligent Information Technology Application 2009:65–68. https://doi.org/10.1109/IITA.2009.390
16. Rashid T, Faizi S, Zafar S. Distance based entropy measure of interval-valued intuitionistic fuzzy sets and its application in multicriteria decision making. *Adv Fuzzy Syst*. 2018;2018:3637897. https://doi.org/10.1155/2018/3637897
17. Scrucca L, Fop M, Murphy TB, et al. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R J*. 2016;8(1):289-317. https://doi.org/10.32614/RJ-2016-021
18. Desgraupes B. Clustering Indices 2017. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. Accessed October 29, 2019.
19. Friedman HP, Rubin J. On some invariant criteria for grouping data. *J Am Stat Assoc*. 1967;62(320):1159-1178. https://doi.org/10.1080/01621459.1967.10500923
20. Grupe G, von Carnap-Bornheim C, Becker C. Rise and fall of a medieval trade Centre: Economic change from Viking Haithabu to medieval Schleswig revealed by stable isotope analysis. *Eur J Archaeol*. 2013;16(1):137-166. https://doi.org/10.1179/1461957112Y.0000000021
21. Hilberg V. Hedeby: an outline of its research history. In: Brink S, ed. *The Viking World*. London, New York: Routledge; 2008:101-111.
22. von Carnap-Bornheim C, Hilberg V. Recent archaeological research in Haithabu. In: Henning J, ed. *Post-Roman Towns, Trade and Settlement in Europe and Byzantium. Vol 1: The Heirs of the Roman West*. Berlin, New York: Walter de Gruyter; 2007:199-218.
23. Jankuhn H. *Haithabu. Ein Handelsplatz der Wikingerzeit*. Karl Wachholtz Verlag: Neumünster; 1986.
24. Jahnke C. … und er verwandelte die blühende Handelsstadt in ein unbedeutendes Dorf. In: Fouquet G, Hansen M, Jahnke C, Schlürmann J, eds. *Von Menschen, Ländern, Meeren. Festschrift für Thomas Riis zum 65. Geburtstag*. Die Rolle Schleswigs im internationalen Handel des 13. Jahrhunderts. Bonn: Stollfuß Verlag Bonn GmbH & Co. Kg; 2006:251-268.
25. Müller U. Haithabu - Schleswig. In: Gläser M, Schneider M, eds. *Lübecker Kolloquium zur Archäologie im Hanseraum X: Vorbesiedlung, Gründung und Entwicklung*. Lübeck: Verlag Schmidt-Römhild; 2016: 339-357.
26. Filzmoser P, Gschwandtner M. mvoutlier: Multivariate outlier detection based on robust methods 2018. http://CRAN.R-project.org/package=mvoutlier. Accessed October 29, 2019.
27. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. *Comput Stat Data Anal*. 2008;52(3):1694-1711. https://doi.org/10.1016/j.csda.2007.05.018
28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018.
29. Desgraupes B. clusterCrit: Clustering Indices 2018. https://CRAN.R-project.org/package=clusterCrit. Accessed October 29, 2019.
30. Kim S. Ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods*. 2015;22(6):665-674. https://doi.org/10.5351/CSAM.2015.22.6.665
31. Göhring A, Mauder M, Vohberger M, et al. Palaeobiodiversity research based on stable isotopes: Correction of the sea spray effect on bone carbonate δ^{13}C and δ^{18}O by Gaussian mixture model clustering. *Palaeogeogr Palaeoclimatol Palaeoecol*. 2018;290:673-686. https://doi.org/10.1016/j.palaeo.2017.11.057
32. Göhring A, Mauder M, Kröger P, et al. Evidence for sea spray effect on oxygen stable isotopes in bone phosphate — Approximation and correction using Gaussian mixture model clustering. *Sci Total Environ*. 2019;673:668-684. https://doi.org/10.1016/j.scitotenv.2019.04.072
33. Bonsall J, Cook G, Lennon R, et al. Stable isotopes, radiocarbon and the Mesolithic–Neolithic transition in the iron gates. *Doc Praehist*. 2000;27:119–132.
34. Clementz A, Gingerich PD, Koch PL. Isotopic records from early whales and sea cows: Contrasting patterns of ecological transition. *J Vertebr Paleontol*. 2006;26(2):355–370. https://doi.org/10.1671/0272-4634(2006)26[355:IRFEWA]2.0.CO;2
35. Clementz MT, Koch PL. Differentiating aquatic mammal habitat and foraging ecology with stable isotopes in tooth enamel. *Oecologia*. 2001;129(3):461–472. https://doi.org/10.2307/4223106
36. Mook WG. Paleotemperatures and chlorinities from stable carbon and oxygen isotopes in shell carbonate. *Palaeogeogr Palaeoclimatol Palaeoecol*. 1971;9(4):245–263. https://doi.org/10.1016/0031-0182(71)90002-2
37. Newsome SD, Martinez del Rio C, Bearhop S, Phillips DL. A niche for isotopic ecology. *Front Ecol Environ*. 2007;5(8):429-436. https://doi.org/10.1890/060150.1
38. Göhring A, Vohberger M, Nehlich O, et al. Approximation of the sea spray effect and limnic influence on δ^{34}S and δ^{15}N values of archaeological human and terrestrial and freshwater animal skeletal finds, In: Grupe G, McGlynn G, Peters J, eds. *Doc Archaeobiol*. Rahden/Westf.: Marie Leidorf GmbH; 2015:169–188.

SUPPORTING INFORMATION

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

How to cite this article: Göhring A, Mauder M, Kröger P, Grupe G. Which isotopes should we choose? Entropy-based feature ranking enables evaluation of the information content of stable isotopes in archaeofaunal material. *Rapid Commun Mass Spectrom*. 2020;34:e8692. https://doi.org/10.1002/rcm.8692