Influence of Oxygen Isotope Ratio on Classification

Markus Mauder, Eirini Ntoutsi, Peer Kröger

May 28, 2014

Abstract

Isotopic fingerprinting is a task of paramount importance for region description and origin prediction. In this work, we use isotopic data, namely oxygen, strontium and lead, from animal remains in the Alps region. Our current samples are not cremated, however the majority of the data to be analysed in the project would be cremated. It is known that oxygen isotopes are not stable under high temperatures, making their application in the analysis of cremated material problematic. We study through Data Mining techniques the effect of oxygen on isotopic fingerprinting (treated as a supervised learning task) and on origin prediction (treated as a supervised task) and explore whether including oxygen in these analyses makes a significant difference to the results.

1 Introduction

Isotopic analysis is used for dating skeletons and archaeological sites, and for diet, climate, and migration patterns [1]. More on general principles and limitations of stable isotope analysis in [5]. The general idea behind these works is that the isotope measurements in the different samples, reflect the environment where these samples were located. Indirectly, this means that different places can be characterized by different isotopic fingerprints or isotopic signatures based on the distribution of the different elements in these places. Extracting the isotopic fingerprint for a place is a task of paramount importance as it provides useful information about the place and it can be also used as a template for classifying new samples to their most probable origin.

In this work, we use samples from animal remaining in the Alps, each of which is described in terms of three elements: oxygen, strontium and lead and their corresponding isotopes.

The goal of the project is to build local isotopic fingerprint models and therefore create different isotopic profiles in the Alps area. Such a profiling would be helpful for comprehending the special characteristics of each area, how the different areas are isotopically-connected and whether the isotopic proximity corresponds to the spatial proximity of the areas. At a second step, this profiling could be used for mapping unknown samples to specific origins, although mapping does not need to rely exclusively on profiling. In Data Mining terms, extracting the isotopic fingerprints is an unsupervised task (clustering) whereas predicting the origins of new samples is a supervised task (classification/ regression).

A critical question at the current state of the project is the small size of the dataset, less than 100 samples which might result in misleading results. Enriching the dataset with new samples would greatly benefit the analysis process. There are further samples which can be employed, however these samples are acquired after cremation. Oxygen is known to be altered at high temperatures and therefore its measurement after cremation is not reliable for our analysis. Except for this "practical issue", the effect of oxygen on isotopic fingerprinting comprises an open discussion in the community. Therefore, in this paper we focus on whether and how the inclusion/exclusion of oxygen affects the results. To this end, we build (using clustering) different fingerprints for the with and without oxygen cases and we compare how the clusters population changes due to the exclusion of

oxygen. We study also the effect of oxygen on the predictive accuracy of classification models.

The rest of the paper is organized as follows: In Sections 2, we focus on oxygen and examine how it is correlated to other isotopes in the dataset and on its spatial distribution in the under investigation area. Next isotopic fingerprinting is presented treated as an unsupervised learning task (clustering) while considering (Sections 3.1) and omitting oxygen (Section 3.2). Predicting the origin of new samples is treated as a supervised learning task (classification) and is discussed in Section 4. Outlier analysis was also performed and its effect on clustering and classification is evaluated (Section 5). A summary and discussion is presented in Section 6. A detailed description of the data, the undertaken preprocessing steps and first exploratory analysis results are given in Appendix A.

2 Exploratory analysis for oxygen

We studied the correlation of oxygen to other isotopes (Section 2) and its spatial distribution (Section 2.2).

2.1 Correlation of oxygen to other isotopes

If all information to be gained from oxygen is apparent from other isotopes, they are either positively or negatively correlated.

The data plotted in Figure 1 indicates that there is no apparent linear correlation between oxygen and the other isotopes. Colors indicate the position of the data point in question in the northern, center, or southern part of the surveyed area (c.f., Section 4). A positive linear correlation would be indicated by the points falling on a diagonal from the origin to the top right, an inverse correlation by a line from top left to bottom right. Since the points in the diagram scatter along no apparent lines there is either a complex (non-linear) correlation or no-correlation at all. If there is indeed no correlation, this may indicate one of two scenarios: either oxygen is orthogonal to the other isotopes and as such highly relevant for differentiation between isotopes, or it is uncorrelated because it does not indicate any association.

2.2 Oxygen distribution in spatial aggregations

The assumption underlying isotope fingerprint analysis is that there is a correlation between samples from the same spatial location. The data set used in this study contains multiple locations represented by more than one sample. To be a viable contribution to the identification of a sample's origin, the distribution of isotopes between locations must be distinct.

In Figure 2, several locations' oxygen isotope distributions are displayed. Some displayed clusters have as little as five points to support their distribution, so the plots can be heavily influenced by outlying values. Despite the low support, it is apparent that the oxygen distribution overlaps between sites.

Figure 3 aggregates regions north (1), inside (2), or south (3) of the Alps. Due to the aggregation of more points the image is more clear. The overlap between regions is an indication that the oxygen isotope is not a strong contributor to the spatial association of isotopes.

3 Building isotopic fingerprints (clustering)

Clustering was employed in order to group the samples into groups of similar samples and extract the isotopic fingerprints per group. For the clustering, we used all isotopic attributes (not the SE attributes though). We used location information later on for the evaluation of the clustering results. Intuitively, we except the extracted groups (without location information) to be spatially distinguishable.

The quality of clustering depends heavily on the quality of the underlying data. A bad isotope will decrease the quality of the model. A good one will bring it closer to the real world. One without any descriptiveness will not influence the model at all.

Here we assume that data were generated by a Gaussian mixture model and we use the well known

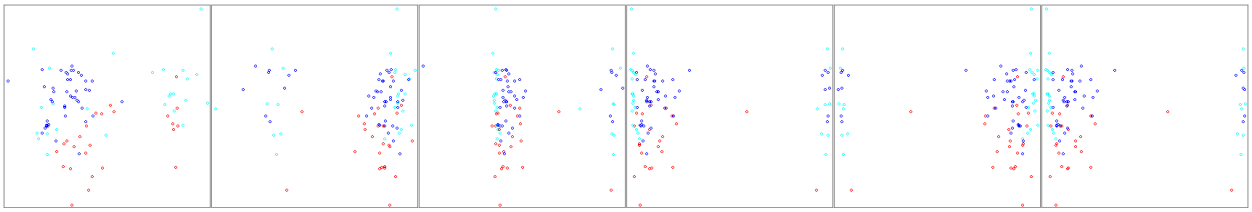


Figure 1: Correlation of oxygen ($18OPO4$, y-axis) with other isotopes. From left to right: $87Sr : 86Sr$, $208Pb : 204Pb$, $207Pb : 204Pb$, $206Pb : 204Pb$, $208Pb : 207Pb$, $206Pb : 207Pb$. Colors correspond to spatial locations: blue (north), red (center), cyan (south) of the Alps.

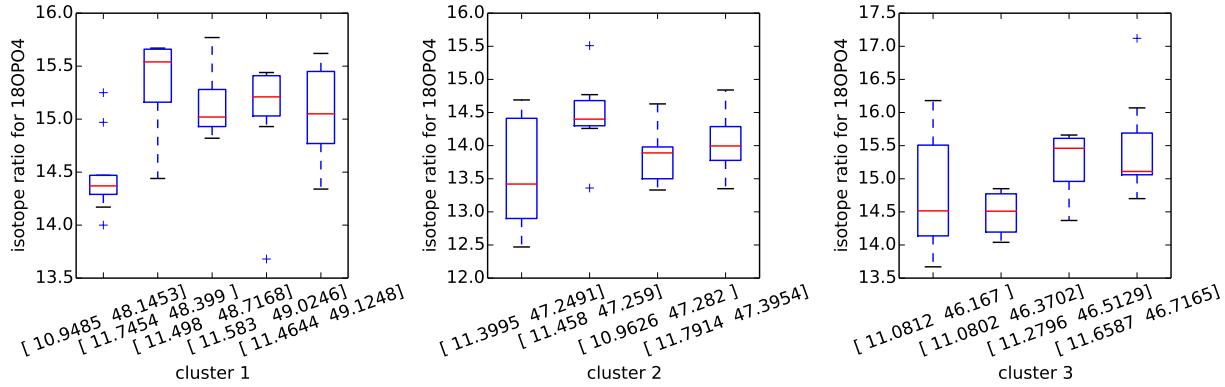


Figure 2: Oxygen isotope distribution by location (cluster 1, 2, 3 corresponds to regions 1 (north), 2 (inside the Alps), 3 (south))

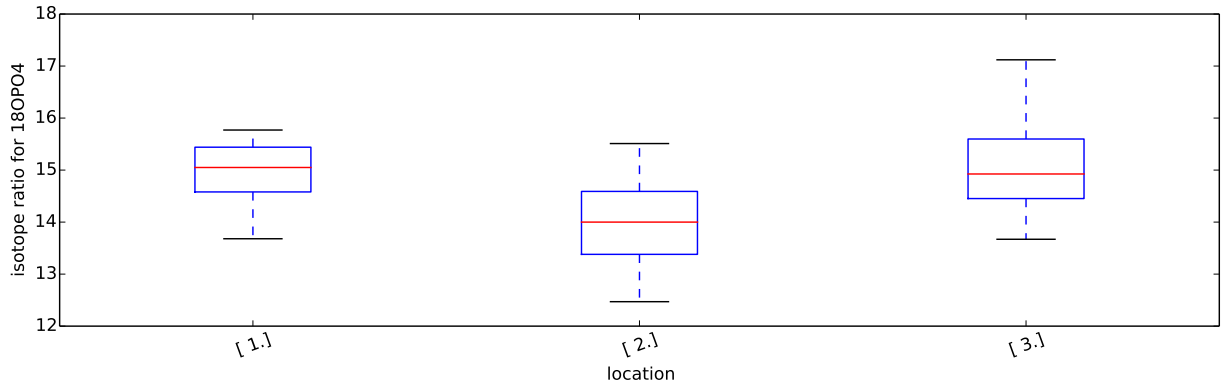


Figure 3: Oxygen isotope distribution by region north (1), inside the Alps(2), south (3).

EM algorithm [2] to estimate the parameters of the model based on our dataset. EM alternates between an expectation E-step that that re-estimates the expected-values of the hidden data (cluster assignments) under the current estimate of the model θ^{old} and the maximization (M) step, which computes new model parameters θ maximizing the expected log-likelihood found on the E step.

After an original guess of k Gaussian distributions parametrized by θ^{old} a different parametrization θ's improvement is calculated by

$$Q(\theta, \theta^{old}) = \sum_{Z} p(Z|X, \theta^{old}) \ln p(X, Z|\theta)$$

where X are samples in isotope space, Z latent variables. The highest improvement is accepted for the next round [2] and the whole process is repeated until convergence.

We used the WEKA implementation of the EM algorithm [3], without specifying the number of clusters to be extracted. The optimal number of clusters is selected by cross-validation. We distinguish between clustering with oxygen (Section 3.1) and without oxygen (Section 3.2). We discuss their differences and effect of oxygen in Section 3.3.

3.1 Considering oxygen

The samples were described in terms of all 7 attributes: 1 oxygen isotope, 1 strontium isotope and 5 lead isotopes.

The algorithm resulted in 6 clusters, the smallest one containing 3 instances, the largest one containing 29 instances. The cluster population is shown in Figure 4.

An overview of the cluster description (mean and standard deviation for each attribute in each cluster) is shown in Figure 5. Mean and standard deviation are not always the best way to display the variation of an attribute values in a cluster, therefore we also provide the boxplots of the different attributes for each cluster. Boxplots are way more informative since except for the mean and standard deviation they also display the median, min, max values, quartiles and

```
Clustered Instances

0      14 ( 15%)
1      29 ( 30%)
2       3 (  3%)
3      16 ( 17%)
4      27 ( 28%)
5       7 (  7%)
```

Figure 4: The population of the extracted clusters (Oxygen case)

Attribute	Cluster 0 (0.07)	1 (0.17)	2 (0.03)	3 (0.17)	4 (0.41)	5 (0.16)
87Sr_86Sr						
mean	0.8339	0.2846	0.3898	0.3439	0.3006	0.8121
std. dev.	0.0167	0.0884	0.1	0.1524	0.1077	0.1512
208Pb_204Pb						
mean	0.7935	0.8214	0.3446	0.2738	0.8724	0.9303
std. dev.	0.0551	0.0638	0.102	0.0885	0.0403	0.0334
207Pb_204Pb						
mean	0.4424	0.4584	0.3591	0.9417	0.396	0.3628
std. dev.	0.0509	0.0293	0.2786	0.0459	0.0225	0.0077
206Pb_204Pb						
mean	0.1525	0.1656	0.32	0.9761	0.0675	0.0108
std. dev.	0.064	0.0371	0.2038	0.0201	0.0293	0.0058
208Pb_207Pb						
mean	0.8132	0.8107	0.6051	0.0241	0.909	0.9795
std. dev.	0.0688	0.0416	0.1853	0.0167	0.031	0.0125
206Pb_207Pb						
mean	0.166	0.1786	0.378	0.9807	0.0749	0.0136
std. dev.	0.066	0.0392	0.1783	0.0179	0.0311	0.0075

Figure 5: Cluster description (Oxygen case)

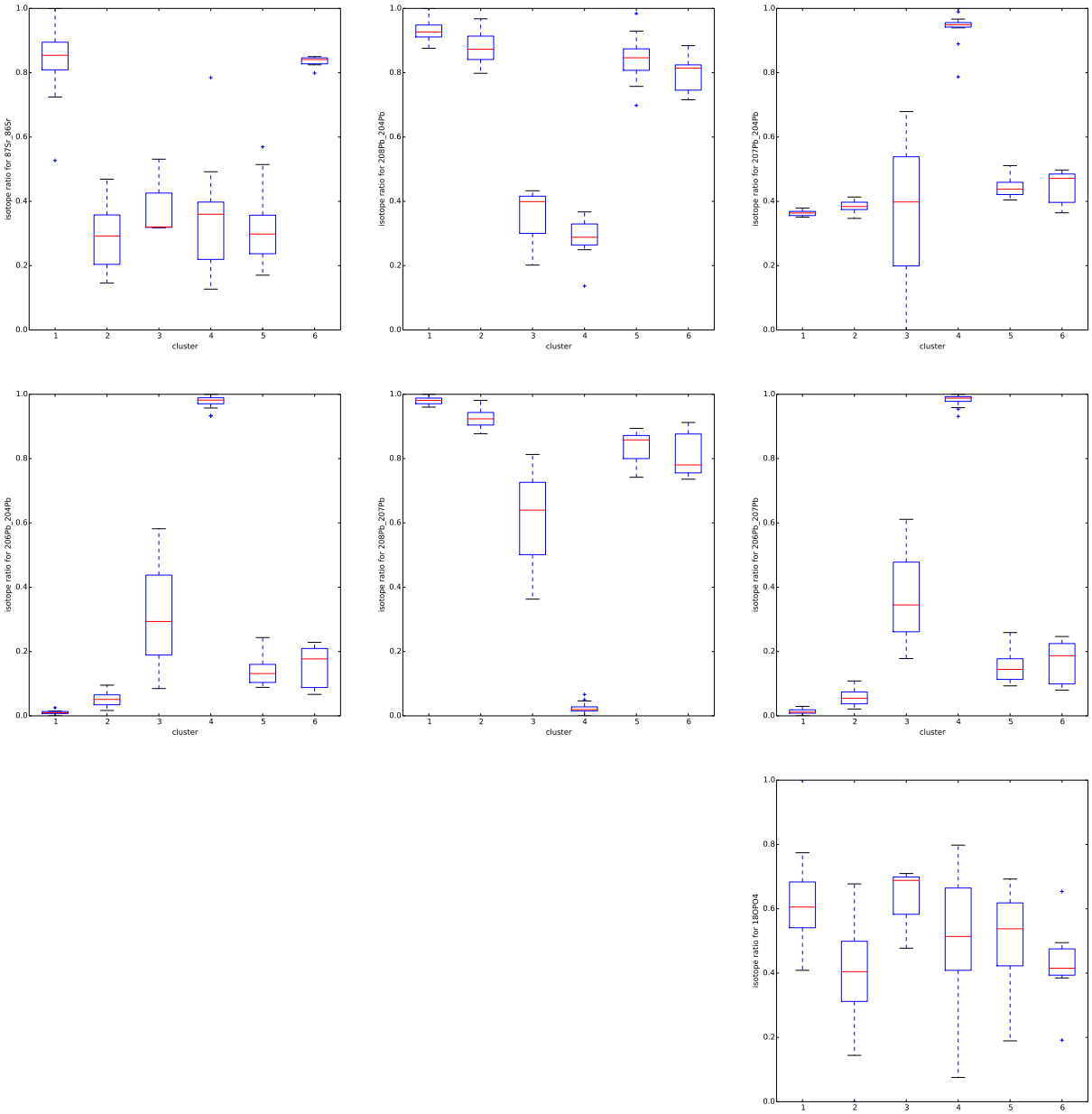


Figure 6: Boxplots per attribute and cluster (Oxygen case)

outliers [1]. The results are shown in Figure 6.

We visualized the clusters versus the samples locations, to show the spatial distribution of the isotope clusters; the results are shown in Figure 7. Each pie in the figure indicates one location. The size of the slices indicates the ratio each of the represented clusters has in that location. Circles of a solid color indicate locations with a single cluster. Spatial distribution of points of a cluster can be estimated by the range of locations containing that cluster's color.

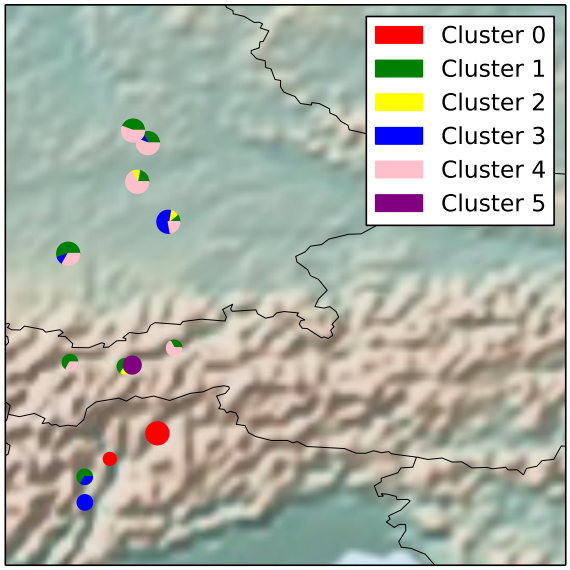


Figure 7: Detected clusters versus locations of the samples (Oxygen case)

We can see that some clusters consists of members which are spatially close like the red (#14 instances) and purple (#7 instances) cluster. However there are clusters scattered over different places like the pink (#27 instances) cluster which is located in Germany-Austria, the green (#29 instances) cluster that is located almost exclusively in Germany-Austria and the blue (#16 instances) cluster that is located in Germany-Switzerland.

[1] http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

So it seems that the resulted clusters based solely on isotopes are not as spatially connected as expected.

3.2 Omitting oxygen

Oxygen is sensitive to cremation procedures, in contrast to strontium and lead elements. Therefore, we want to check how the clustering results, and therefore the isoscaping, is affected by excluding oxygen from clustering. To this end, we repeat the clustering experiment but this time we leave out the oxygen attribute and we rely solely on strontium and lead isotopic values.

The new clustering results also in 6 clusters, whose population is displayed in Figure 8.

```
Clustered Instances

0       7 (  7%)
1      15 ( 16%)
2       3 (  3%)
3      16 ( 17%)
4      40 ( 42%)
5      15 ( 16%)
```

Figure 8: The population of the extracted clusters (No-oxygen case).

The actual cluster description (mean and standard deviation for each cluster) is shown in Figure 9. For a better comparison across the different clusters, we also display the boxplots for each cluster along each dimension, in Figure 10.

We visualized the clusters versus the samples locations, the results are shown in Figure 11. The picture is quite similar to what we observed for the non-oxygen case (c.f., Figure 7). Similarly to the oxygen case clustering, only two clusters consists of members located in the same area, namely the pink and the cyan cluster. The rest are spread across different places, for example the black and white clusters are located all over the three countries, whereas the orange cluster is located in Germany and Austria.

3.3 The effect of oxygen on clustering

A visual inspection of the with-oxygen (Figure 7) and without-oxygen (Figure 11) clustering indicates that

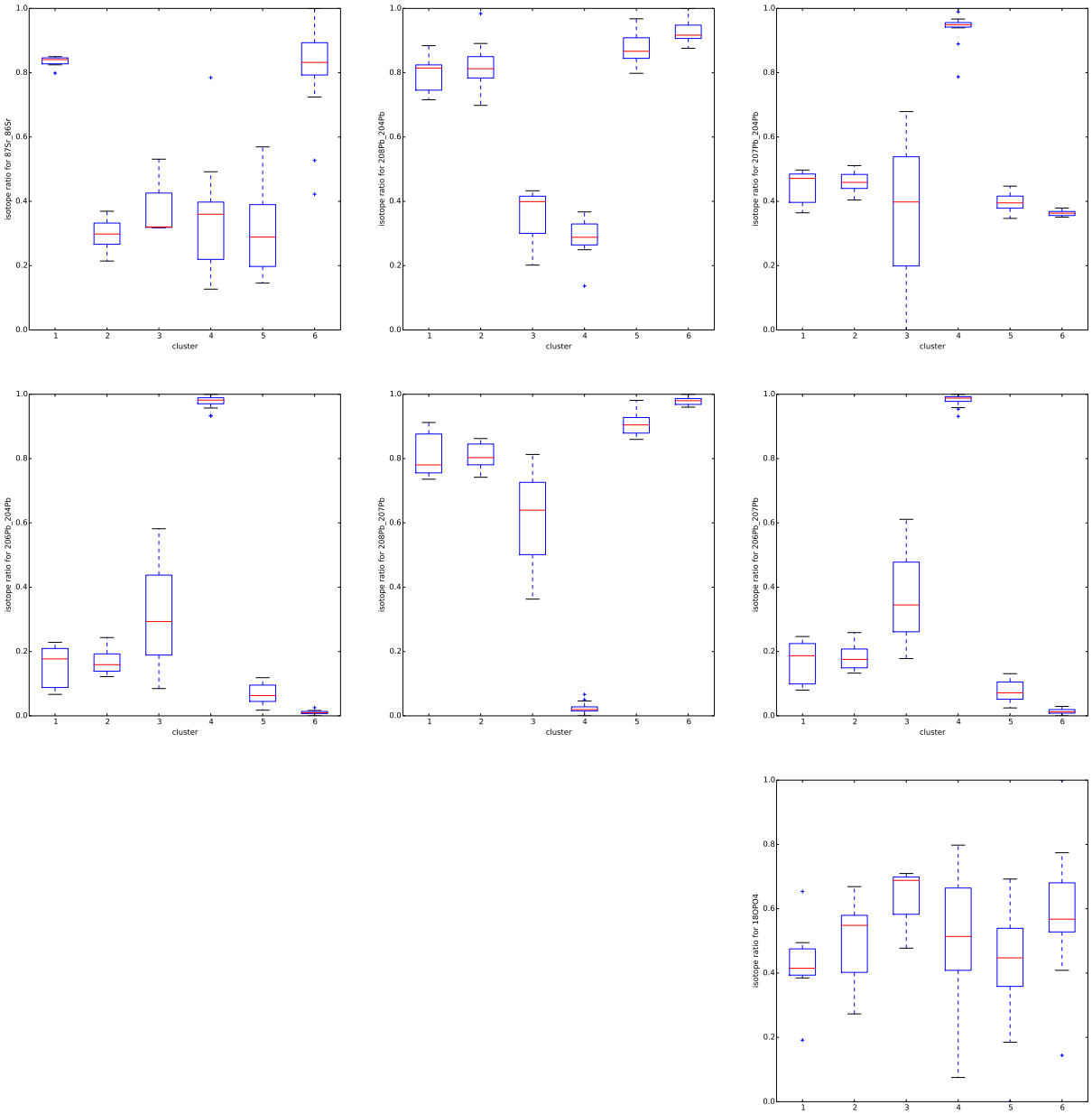


Figure 10: Boxplots per attribute and cluster (No-oxygen case).

Attribute	Cluster 0 (0.07)	1 (0.16)	2 (0.42)	3 (0.18)	4 (0.17)
87Sr_86Sr					
mean	0.7165	0.7162	0.7105	0.7111	0.7103
std. dev.	0.0002	0.0017	0.0012	0.0017	0.001
208Pb_204Pb					
mean	38.2617	38.6358	38.4757	36.8658	38.2624
std. dev.	0.1508	0.0914	0.11	0.2563	0.3279
207Pb_204Pb					
mean	15.8198	15.6403	15.7166	16.9102	15.8515
std. dev.	0.1148	0.0174	0.0512	0.1717	0.0716
206Pb_204Pb					
mean	21.0094	18.6921	19.6395	34.0993	21.3977
std. dev.	1.0467	0.0941	0.49	1.5507	0.7584
208Pb_207Pb					
mean	2.4187	2.4703	2.4481	2.1805	2.4137
std. dev.	0.0213	0.0039	0.0098	0.0252	0.0176
206Pb_207Pb					
mean	1.3274	1.1951	1.2495	2.016	1.3498
std. dev.	0.0573	0.0065	0.0277	0.077	0.0471

Figure 9: Cluster description (no-oxygen case).

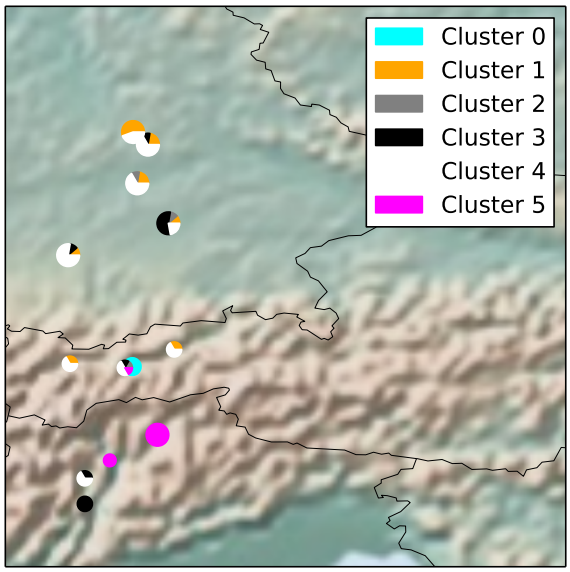


Figure 11: Detected clusters versus locations of the samples (No-oxygen case).

some samples are co-clustered in both cases. However, in order to understand the exact mapping of the clusters from the oxygen to the non-oxygen clustering case, we relied upon the intersection of the cluster members following the MONIC framework [6].

As we can see, cluster 0 of the oxygen case (red color) *survived* entirely to cluster 5 of the non-oxygen case (pink color). Also, cluster 2 of the oxygen case (yellow color) *survived* entirely to cluster 2 of the non-oxygen case (gray color). Similarly, cluster 3 of the oxygen case (blue color) *survived* entirely to cluster 3 of the non-oxygen case (black color). Moreover, cluster 5 of the oxygen case (purple color) *survived* entirely to cluster 0 of the non-oxygen case (cyan color). Cluster 1 (green color) almost exclusively *survived* into cluster 4 (white color) and a tiny percentage to cluster 5 (pink color) A split occurred, namely cluster 4 of the oxygen case (pink color) was *split* into two clusters for the non oxygen case, cluster 1 (orange color) and cluster 4 (white color). Cluster 4 of the non-oxygen case (white color) is the result of *merge* from cluster 1 (green color) and cluster 4 (pink color) of the oxygen case.

So, in total some clusters are entirely untouched by the omission of oxygen, whereas two others were involved in merge and splits operations. The clusters merging and splitting are on the one hand spatially close and on the other they comprise a large region, indicating that they are very broad themselves. This explains some of the instability of these clusters and that they are susceptible to disruption by not very indicative isotopes. These findings supports the notion that oxygen is not a key feature for isotopic fingerprinting.

4 Supervised analysis: classification with and without oxygen

From the spatial information associated with samples, a class label can be derived, which can then be used to build a model for this set of samples. More specifically, we categorized the data based on their spatial coordinates into classes “north”, “mid-

		no-oxygen clustering					
		cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Oxygen clustering	cluster 0	0	0	0	0	0	1.0
	cluster 1	0	0	0	0	0.97	0.03
	cluster 2	0	0	1.0	0	0	0
	cluster 3	0	0	0	1.0	0	0
	cluster 4	0	0.56	0	0	0.44	0
	cluster 5	1.0	0	0	0	0	0

Table 1: Migration of samples between clusters when using oxygen vs no oxygen isotope (cluster names are automatically generated by Weka).

dle" and "south" Alps. This is termed a "supervised" data mining task. Based on these models, the association with a previously unseen sample can be established.

A classification model is built upon a training set of known class labels and its performance is evaluated over a test set of know labels which are employed during training. The idea is that the model should be able to best describe the training set but also to generalize in case of unseen instances by the model.

To judge the effect of oxygen on classification, we build two classification models, one considering and one omitting oxygen, and we compare the classifiers performance. If the classification performance grows with the omission of oxygen, its inclusion had a detrimental effect on the quality of the classification model. If it shrinks oxygen contributes to the performance. Finally, if the classification performance is not significantly affected, the omission of oxygen has no effect.

Table 2 shows a few indicating values about the quality of the classification with oxygen and without it.

A summarizing value that can be used for direct comparison is the F-Measure. It is defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The reported values are:

TP: True Positive: hits

FP: False Positive: misses

True Positive Rate: $TP/(TP + FN)$

False Positive Rate: $TP/(TP + FP)$

Precision pro Klasse: $TP/(TP + FP)$

Recall pro Klasse: $TP/(TP + FN)$ average == TP Rate

F-Measure (also *F1 score*): $2 \cdot Prec \cdot Recl/(Prec + Recl)$

ROC Area *Receiver Operating Characteristic Area under Curve*

In direct comparison, classification values with oxygen are a bit higher than the ones without oxygen, but not by a large margin (0.83 vs 0.76). This seems to indicate that oxygen can indeed contribute to the classification result.

4.1 Univariate classification based on single isotopes

Table 3 shows the classification performance based on distinct isotopes. The performance of the univariate classifiers built upon single isotopes is very low and much lower comparing to the multi-variate classifier (c.f., Section 4). Surprisingly, oxygen performed best of all isotopes tested in this experiment. This may indicate that it is helpful, but stands in direct opposition with the findings of Section 4, where its removal had relatively little effect. It is important to note that there is no direct correlation between an attribute's individual classification power and the aggregation of multiple attributes. For an example, compare the f-measure with the omission of oxygen at 0.759 to that

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Oxygen	0.833	0.115	0.837	0.833	0.832	0.868
No Oxygen	0.76	0.168	0.768	0.76	0.759	0.785

Table 2: Classification quality.

of strontium at 0.674. Strontium has more influence on the aggregated performance than oxygen although it performs worse in the individual classification.

5 Outlier detection and their effect on the analysis

According to Hawkins [4] definition "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism". Outliers have a negative effect on data mining tasks and therefore, it is important to see whether there exist outliers in our dataset. For the outlier detection, we rely on the interquartile range test and consider as *extreme outliers* all those points that belong to the lower outer fence (Q1 - 3*IQ) and the upper outer fence (Q3 + 3*IQ). IQ is the interquartile range (Q3 - Q1), where Q1 is the lower quartile (the 25th percentile) and Q3 is the upper quartile (the 75th percentile). A visual explanation is given in Figure 12, where the extreme outliers area is pointed out in red.

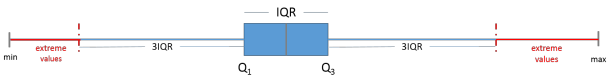


Figure 12: Extreme outliers test

Instances which contain at least one attribute with outlier values are considered as outlier instances. The outlier analysis resulted in 17 instances containing outlier values out of the 96 instances of our original dataset.

The outlier versus the non outlier instances were spatially projected and are depicted in Figure 13. One can see that the outliers (green color) are located in the south and in the north. A closer inspection

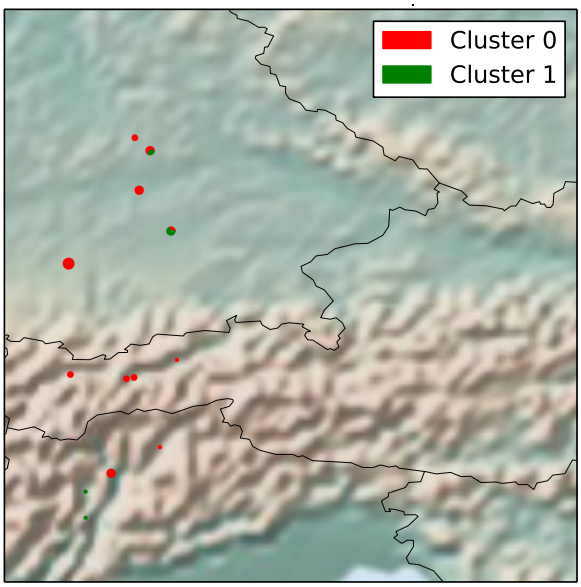


Figure 13: Outlier instances (green color) vs non-outlier instances (red color).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
18OPO4	0.542	0.297	0.536	0.542	0.531	0.665
87Sr	0.51	0.252	0.524	0.51	0.515	0.645
208Pb	0.479	0.302	0.48	0.479	0.479	0.576

Table 3: Clustering using only one isotope

of the results and an association with the clustering results presented in the previous sections shows that the outlier instances are mainly members of a single cluster (blue cluster for the oxygen case and black cluster for the non-oxygen case). This cluster seems quite "suspicious" spatially in the sense that its members reside in north or south and there are no members in the middle Alps. This is in contrast to other spatially mixed clusters that are spread all over the three countries.

We filtered out the outlier instances and build a classifier model (kNN classifier, k=1) over the cleaned instances (#79 instances). The performance of the classifier is displayed in Table 4. In average (last row of Table 4), there is an improvement of 5% compared to considering the whole dataset case which also includes the outlier points.

6 Discussion and Conclusions

The number of data points analyzed is quite small (#96 data points) and scattered over only few sites (#13). This is particularly problematic when the goal is to build a model of the covered areas (isotopic fingerprinting) and use these models for origin prediction for future samples. To make things more severe, some of the points stand out as different (outliers). If these points were spatially constrained, this would indicate a particularly clear fingerprint for that region. However, the outlier points are not constrained, rather they are scattered all over the covered area, allowing no such conclusion. Ignoring the outliers from the analysis, results in further shrinkage of the training set and therefore the danger of over-fitting is even larger.

Based on these limitations, we took a first step towards analyzing these data in terms of clustering, classification and outlier detection for isotopic fingerprinting and origin prediction. We focused in this study on the effect of oxygen, which is sensitive to high temperatures, and how its inclusion/ exclusion affects the results of the corresponding analysis. Our findings in terms of extracted cluster and classification models with and without oxygen indicate that oxygen does not contribute strongly to the results.

A Data description

A.1 Data preprocessing

The original dataset consists of 99 samples. Each sample is described in terms of isotopes of three elements: Oxygen (^{18}O), Strontium (^{88}Sr, ^{87}Sr, ^{86}Sr), and Lead (^{204}Pb, ^{206}Pb, ^{207}Pb, ^{208}Pb) , as well as descriptive attributes like animal species type and skeleton part. Four samples to be remeasured were omitted. Attributes *XRD*, *ID OssoBook*, δ^{18} *O co2 SMOW hofes* were omitted as irrelevant for the analysis. The attribute *Labor Nr.* was used to extract the spatial coordinates of the samples, so each sample was enhanced by *latitude* and *longitude* attributes. The standard error columns associated with each isotopic measurement where also removed at this phase, but will be incorporated in the future. The final dataset consists of 96 samples. Two species types, "hirsch" and "rothirsch" were merged into a single spezie type, "hirsch".

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1 (North)	0.946	0.143	0.854	0.946	0.897	0.91
2 (inner Alpine)	0.75	0.018	0.947	0.75	0.837	0.878
3 (South)	0.944	0.033	0.895	0.944	0.919	0.96
weighted avg	0.886	0.08	0.891	0.886	0.884	0.912

Table 4: Performance of classifier on cleaned data set.

Attribute name	**Description**
18OPO4	Oxygen isotope
87Sr_86Sr	Strondium isotope
208Pb_204Pb	Lead isotope 1
207Pb_204Pb	Lead isotope 2
206Pb_204Pb	Lead isotope 3
208Pb_207Pb	Lead isotope 4
206Pb_207Pb	Lead isotope 5

A.2 Exploratory analysis of the data

The main characteristics of the data were evaluated through exploratory data analysis. We focused on the following aspects of the data: i) class distribution of the samples, where class is the species type (Section A.2.1), ii) spatial distribution of the samples (Section A.2.2), iii) skeleton part distribution of the samples (Section A.2.3), iv) correlations between attributes (Section A.2.4). For each aspect, we provide results hereafter.

A.2.1 Class distribution

As class we considered the species types, the distribution is shown in Figure 14. Classes "rind", "schwein" are almost equally represented, whereas class "hirsch" is slightly underepresented.

The class distribution for the different attributes is shown in Figure 15. We can see that the attributes follow different distributions, but different classes are present at almost every value range.

A.2.2 Spatial distribution

The samples' locations was projected in Google maps, the results are shown in Figure 16. In the original proposal, more locations were described, as shown in Figure 17. By juxtaposing these locations to the

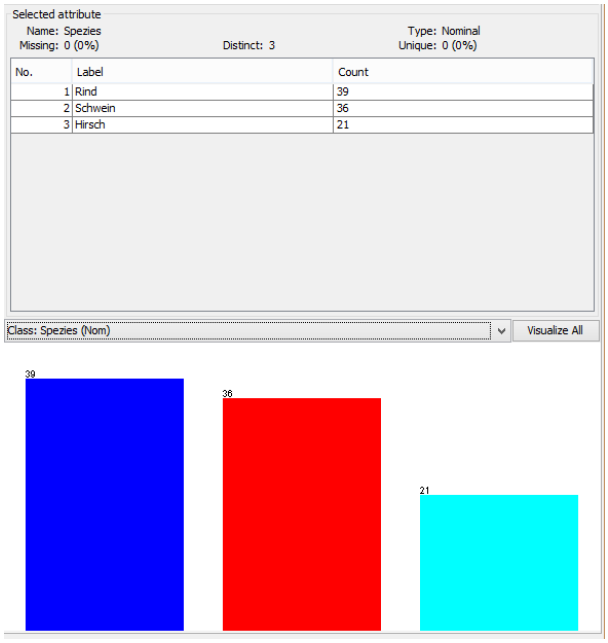


Figure 14: Class distribution of the samples

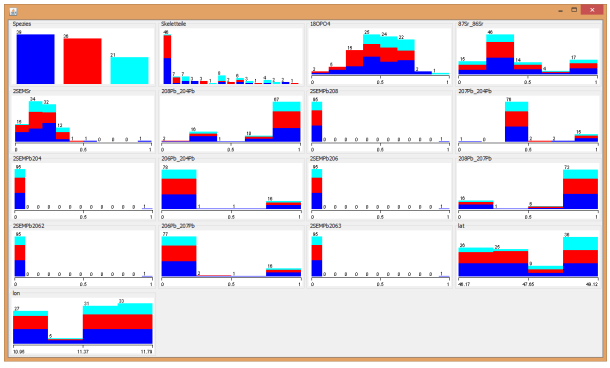


Figure 15: Class distribution for the different attributes

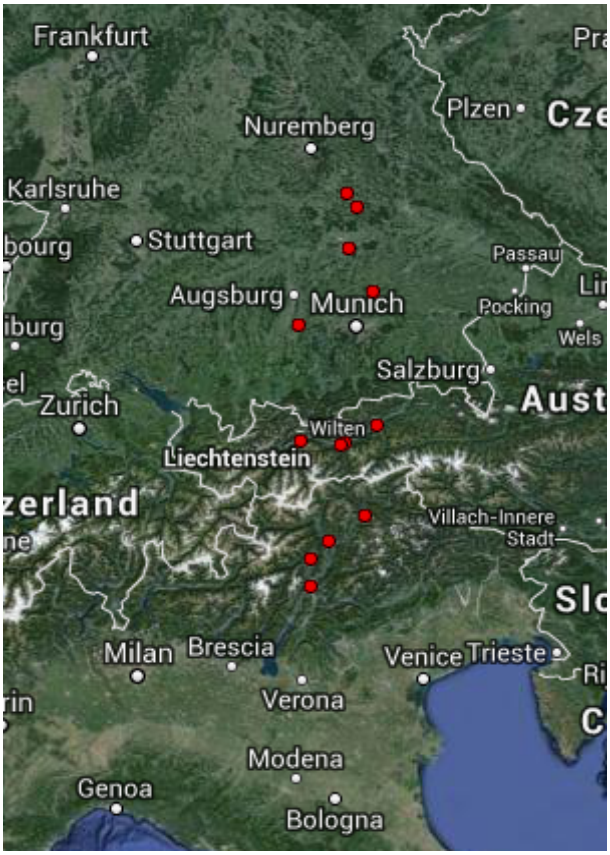


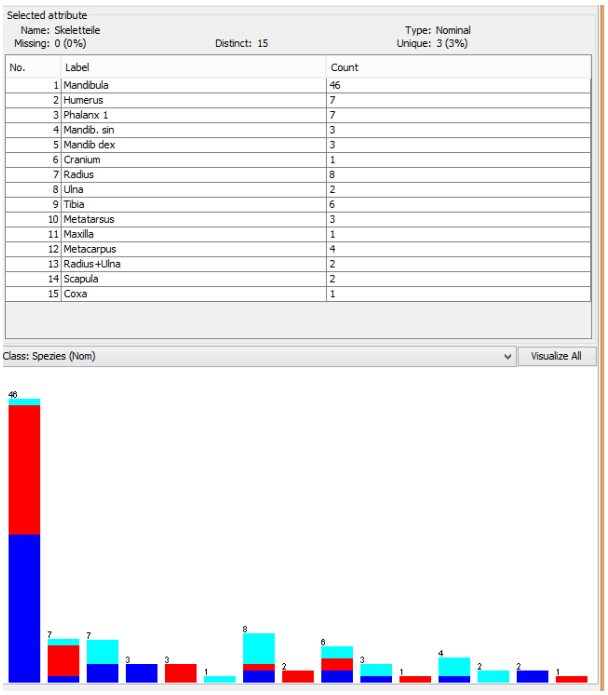
Figure 16: Spatial distribution of the samples

Figure 17: Spatial coverage in the original proposal

samples' locations map (Figure 16), we can see that the sample coverage of the area under study is incomplete, most archaeological sites are not present. Also, the majority of the samples comes from Germany, then Austria and finally Switzerland.

A.2.3 Skeleton parts distribution

The samples come from different parts of the animals bodies, in Figure 18 we should the distribution of the different parts, as well as the class distribution within each distinct skeleton part.



No.	Label	Count
1	Mandibula	46
2	Humerus	7
3	Phalanx 1	7
4	Mandib. sin	3
5	Mandib dex	3
6	Cranium	1
7	Radius	8
8	Ulna	2
9	Tibia	6
10	Metatarsus	3
11	Maxilla	1
12	Metacarpus	4
13	Radius+Ulna	2
14	Scapula	2
15	Coxa	1

Figure 18: Skeleton distribution of the samples and within species type distribution (blue: rind, red: schwein, cyan: hirsch)

One can see that most of the samples come from the "mandibula" skeleton part and there are 15 distinct samples skeleton parts. We make the assumption that the skeleton part used for sampling does not have any impact on the results, besides the selection of the skeleton part for sampling is driven by the actual findings, researchers in this field use mainly the "mandibula" part but if this is not available other parts of the skeleton are employed.

A.2.4 Attribute correlation

An overview of the correlations between the different attributes is shown in Figure 19.

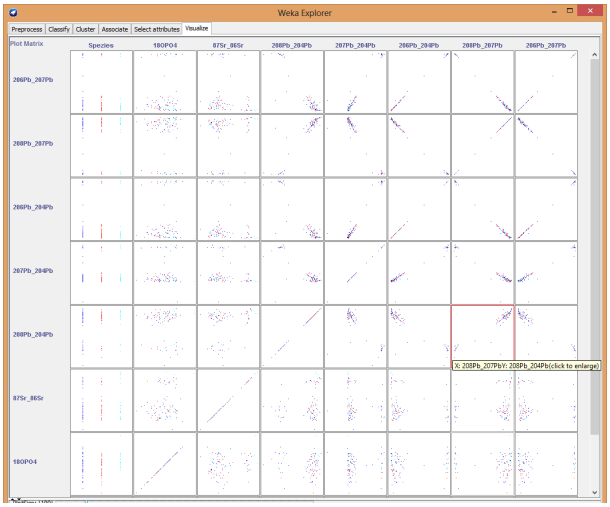


Figure 19: Correlations between attributes

One can see that there are positively correlated attributes and negative correlated ones. Positively correlated attributes: ($^{206}Pb/^{207}Pb$, $^{207}Pb/^{204}Pb$); ($^{206}Pb/^{207}Pb$, $^{206}Pb/^{204}Pb$); ($^{208}Pb/^{207}Pb$, $^{208}Pb/^{204}Pb$);

Negatively correlated attributes: ($^{206}Pb/^{207}Pb$, $^{208}Pb/^{204}Pb$); ($^{206}Pb/^{207}Pb$, $^{208}Pb/^{207}Pb$); ($^{208}Pb/^{207}Pb$, $^{207}Pb/^{204}Pb$); ($^{208}Pb/^{207}Pb$, $^{206}Pb/^{204}Pb$);

So, positive/negative correlations exist mainly among isotopes of lead. There are attributes that seem to be non linearly correlated, like Oxygen isotopes versus Strontium or Lead isotopes.

References

[1] S. H. Ambrose and J. Krigbaum. Bone chemistry and bioarchaeology. *Journal of Anthropological*

Archaeology, 22(3):193 – 199, 2003. Bone Chemistry and Bioarchaeology.

[2] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. Springer New York, 2006.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[4] D. Hawkins. *Identification of Outliers*, volume 1. Chapman and Hall, 1980.

[5] W. Meier-Augenstein and H. F. Kemp. *Stable Isotope Analysis: General Principles and Limitations*. John Wiley and Sons, Ltd, 2009.

[6] M. Spiliopoulou, E. Ntoutsi, Y. Theodoridis, and R. Schult. MONIC: modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA*, pages 706–711, 2006.