Applying Data Mining Methods for the Analysis of Stable Isotope Data in Bioarchaeology

Markus Mauder*, Eirini Ntoutsi†, Peer Kröger*, Christoph Mayr‡, Gisela Grupe§, Anita Toncala§, and Stefan Hölzl¶
*Institute for Informatics, Data Science Lab, Ludwig-Maximilians-Universität München, Germany
Email: {mauder,kroeger}@dbs.ifi.lmu.de
†Faculty of Electrical Engineering and Computer Science, Leibniz Universität Hannover, Germany
Email: ntoutsi@kbs.uni-hannover.de
‡Institute for Geography, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
Email: c.mayr@lrz.uni-muenchen.de
§Bio-Center, Ludwig-Maximilians-Universität München, Germany
Email: {g.grupe,anita.toncala}@lrz.uni-muenchen.de
¶RiesKraterMuseum Nördlingen, Germany
Email: s.h@lmu.de

***Abstract*—Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is choosing good features for analysis. In this paper, we introduce a data science framework that was designed to allow domain experts to consider their domain knowledge in assembling suitable data sources for complex analyses. The structure of experimental data as represented by a clustering is used to measure the relevance as well as the redundancy of each feature. We present an application of this technique to bioarchaelogical data from a region in the European Alps, a transalpine passage of eminent archaeological importance in European prehistory, the Inn-Eisack-Adige passage, spanning Italy, Austria, and Germany. These results are applied to the task of provenance analysis. The application of the presented data mining technique leads to new insights which were not found using standard bioarchaeological approaches.**

I. Introduction

Multivariate evaluation of measurement data has become common practice in many sciences. However, faced with multi-dimensional data many domain scientists struggle to understand analysis results. A common task in multi-variate data analysis is assessing which features to generate and consider in the first place. While there are fully automatic multi-variate feature analysis methods, they do not allow the domain scientist to evaluate a feature's merit in light of domain-specific considerations. In this paper we introduce a feature evaluation technique that allows domain scientists to understand each feature's role in the data distribution and to evaluate its importance for the analysis at hand.

The project which motivated this paper aims at the construction of a large scale isotopic map covering a specific transect across the European Alps, namely the Inn-Eisack-Adige passage via the Brenner pass. This transect has been in use at least since the Mesolithic and is therefore of eminent archaeological importance. The isotopic mapping of the transect aims at answering open archaeological questions related to transalpine mobility and culture transfer[1]. The term *isotopic landscape* describes *maps of isotopic variation produced by iteratively applying (predictive) models across regions of space using gridded environmental data sets*, whereby one *common use of isoscapes is as a source of estimated isotopic values at unmonitored sites, which can be an important implementation for both local- and global-scale studies if the isoscape is based on a robust and well-studied model* [1]. In bioarchaeology, variations of isotopes (atoms with different numbers of neutrons) are used to predict patterns that characterize the origin of geological and biological materials at a small spatial scale. Such isotopic maps are empirically generated by sampling the relevant environmental components and measuring their isotopic signatures. However, the vast majority of stable isotope studies in this field are small scale projects that lack the fundamental capabilities of prediction and modeling.

In this paper, we use isotope data to investigate the question which features should be measured in order to keep the costs for generating a reliable data source for this isotope map moderate. We describe a framework that was developed to solve this problem and, thus, supports the domain experts in making decisions during data generation. The data mining task was to discover which features were most relevant or most redundant and therefore irrelevant for analysis. The results are of high value for the domain scientists. Our framework for solving the aforementioned data mining task is based on a technique that explores the relevance and redundancy of individual variables to a clustering in comparison to a reference clustering. There is no obvious reference clustering, because no ground truth is known. However, domain specific knowledge and assumptions can be used to generate several *plausible* reference clusterings that are estimations of a ground truth. Thus, we explore how the relevance and the redundancy of single features behave under different ground truth

[1]See www.for1670-transalpine.uni-muenchen.de

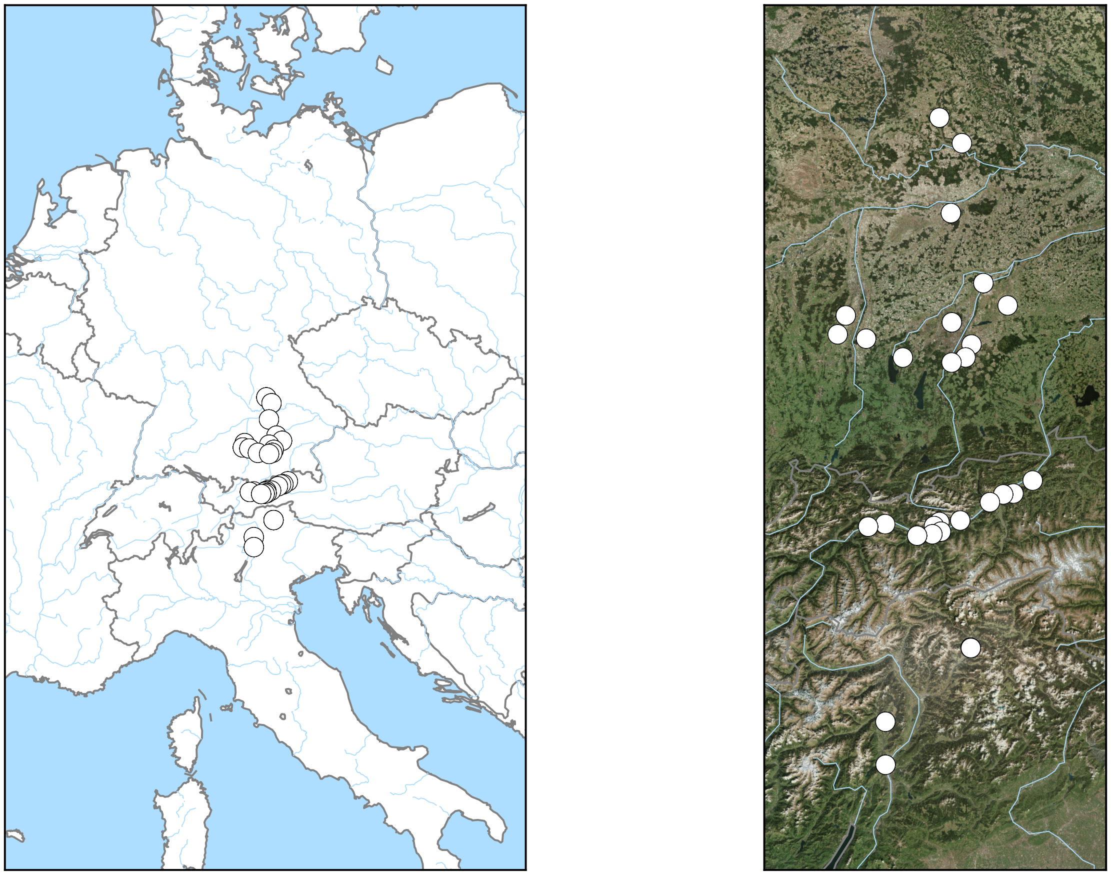


Fig. 1: Sampling sites across the transalpine Inn-Eisack-Adige passage. Data from these locations is used in the evaluation in Section III.

assumptions and derive conclusions from these observations. The results have been carefully discussed with the domain experts and the resulting conclusions confirm that we were able to benefit their work.

In summary, our contributions are as follows:

- We describe a new framework that solves a real problem in the application domain. The framework supports identifying relevant features that need to be measured and redundant features that need not to be measured. The framework is a real data science product, because it emerged in tight interdisciplinary collaboration.
- We describe a new application from archaeobiology that strongly benefits in multiple ways from an interdisciplinary data science approach.
- Based on problems in the application domain, we identify new challenges for the data mining community that will enable researchers from archaeobiology to improve experiment design and understand the resulting data.
- We discuss the resulting conclusions and added values for the domain experts.

The rest of the paper is organized as follows: In Section II we describe our data-driven approach for the analysis and the evaluation of individual features. In Section III we introduce the task of isotopic mapping, the underlying data, and important challenges. We also present our dataset, which comes from a specific region in the Alps in Europe, experimental results, evaluated by domain experts, and insights gained through this study. In Section IV we conclude our study and highlight challenges and opportunities in this domain for the data mining community.

II. Feature evaluation

In our approach, the quality of a feature, or feature subset, is assessed based on its contribution to one ore more reference data structures. In particular, we assess how stable (i.e., unchanged) the data structure is across feature space projections. Our assumption is that a highly relevant projection will result in a data structure that resembles the reference data structure.

The task of assessing the importance of a feature for provenance analysis is reminiscent of feature selection and feature ranking. Feature selection generates a subset of the most suitable features for a given task, whereas feature ranking returns an ordering of features according to their importance for the task [2]. Most of the common approaches are supervised, meaning that they require class labels for accessing the quality of a feature or feature subspace [3]. Such information is not available for the discussed usecase, therefore we have to rely on unsupervised feature selection approaches [4]. In particular, we follow a wrapper-based approach [5] where we use a learning algorithm (EM clustering in our case) for the evaluation of a feature or a subspace. A big part of research on feature selection and feature ranking methods is focused on reducing the exponential search space of possible solutions. In our case, the feature space is low-dimensional but the domain scientists are interested especially on understanding i) the importance of each feature for the final model and ii) whether there are other features in the feature space that can replicate the "contribution" of that feature. The reason is that feature acquisition is an expensive process as domain experts have to follow lengthy and time consuming processes of cleaning the findings and measuring the isotope values. Moreover, in some cases it is not possible to measure all differenet isotopes for all available samples. This is the case for our project, where the oxygen isotope cannot be measured for cremated human findings. So it is extremely important for the domain experts to understand whether oxygen is a key feature for the analysis and also whether the remaining features can compensate for oxygen's contribution to the final model. Therefore, we follow a clustering-based feature evaluation approach, where we compare unsupervised learning results that convey aspects of the data structure (from a single feature point of view) with the data structure (as captured by the reference clustering). That is, the data structure is represented by the clusters extracted from the data. To assess the effect of a projection on the data structure, we compare the projection-based partitioning to the reference data partitioning.

Our proposed unsupervised feature evaluation framework consists of three steps:

1) data structure extraction (clustering)
2) data structure comparison (Adjusted Rand Index)
3) feature evaluation

Before explaining each of these steps, we introduce some notation: Let D be a dataset in a feature space F. Let $F_0 \subset F$ be the feature set from which a model of the reference data structure is extracted by clustering; we refer to Θ^{F_0} as the *reference clustering* and to F_0 as the *reference feature space*. Let $F_v \subset F$ be a set of features to investigate w.r.t. their quality for the reference data structure, Θ^{F_0}. Note that F_v and F_0 are treated as being independent from each other even though they need not be disjunct.

A. Unsupervised data structure extraction

To extract structure from the data, we use a clustering approach [6]. Domain knowledge suggests continuous values for the measurements, which can be best modeled as a mixture model of continuous distributions, like a Gaussian Mixture Model. To extract a robust indication of the data's structure in an unsupervised way, we applied the Expectation-Maximization (EM) algorithm [7]. EM fits a number of multi-variate normal distributions over the given data set. The result is a soft-clustering; in our dataset though the assignment is typically fairly hard. A typical run over the isotope dataset (see Section III-C) results in a standard deviation of 0.115

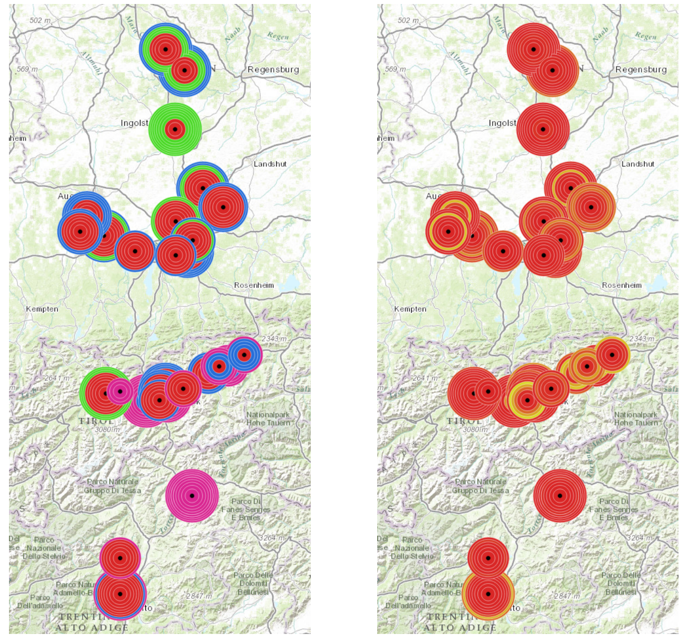


Fig. 2: Example EM clustering on isotope data. Image best viewed in color. Left: spatial projection of hard clustering (maximum likelihood cluster). Right: membership likelihood of soft clustering (red highest, yellow lowest).

for the maximum likelihood cluster labels. As an example, Figure 2 depicts the membership and spatial structure of a sample clustering done on the set of 217 bones where only the seven isotopic features are used. For easier handling, we convert the cluster probabilities to hard cluster assignment by their maximum likelihood. The result of the clustering is a set of partitions, $\Theta^F = \{\theta_1, \theta_2, \ldots, \theta_k\}$, where k is the number of clusters (optimized by cross-validation, see below).

B. Comparing clusterings

To compare how well a clustering Θ^{F_v} extracted from an investigated feature projection F_v reflects the structure of a reference clustering Θ^{F_0}, we employ the *Adjusted Rand Index* (ARI) [8] of the two clustering partitionings:

$$s(F_0, F_v) := ARI(\Theta^{F_0}, \Theta^{F_v})$$

ARI evaluates the agreement between two clusterings by counting pairs assigned to the same cluster under both clusterings and pairs assigned to different clusters versus the total number of pairs in the dataset. ARI was proposed to reduce the influence of randomness on the traditional Rand Index (RI) [9] and has been proven to perform better when the number of clusters in the two clusterings is not the same [10], [11]. Like the rand index, ARI has a maximum value of 1 and takes the value 0 when the index equals its expected value. However, negative values are also possible and indicate an agreement that is even less than one expected between two random clusterings.

C. Unsupervised feature evaluation

Not all attributes are equally important for a given analysis task: A feature may be unnecessary to describe the result of a given analysis or the data reflected in the feature may be noise or encompassed by other attributes. By selecting a suitable comparison feature space we investigate the *structural relevance* of a feature (i.e., how well it captures the structure in isolation) for a clustering as well as its *structural redundancy* (i.e., if the clustering becomes unstable without this particular feature).

To generate these scores, we extract a single feature $f \in F_v$. Let D_f be our original dataset projected onto dimension f and let Θ^f be the clustering over D_f : $\Theta^f = \{\theta_1, \theta_2, \dots, \theta_{k'}\}$, where k' is the number of clusters. We refer to Θ^f as the *univariate clustering*. Let $f_- = F_v \setminus f$ be the complementary feature space, that is, all dimensions in F_v except for the investigated feature f. Let D^{f_-} be the complementary dataset, i.e., the dataset projected onto the complementary feature space f_-. Applying EM on D^{f_-} generates a clustering $\Theta^{f_-} = \{\theta_1, \theta_2, \dots, \theta_{k''}\}$ where k'' is the number of clusters. We refer to Θ^{f_-} as the *complementary clustering*.

To calculate the structural relevance of f, we compare the univariate clustering Θ^f derived from the specific feature f to the reference clustering Θ^{F_0}:

$$s_{relevance}(f, F_0) := ARI(\Theta^f, \Theta^{F_0})$$

To calculate the structural redundancy of f, we compare the complementary clustering Θ^{f_-} derived from the complementary feature space f_- to the reference clustering Θ^{F_0}:

$$s_{redundancy}(f, F_0) := ARI(\Theta^{f_-}, \Theta^{F_0})$$

The first comparison evaluates the structural relevance of f for Θ^{F_0}, whereas the second evaluates whether f's contribution can be reproduced by other features in the feature space. In that sense, the first score derives the specific feature's structural relevance and the second score its structural redundancy due to the existence of other feature(s) in the feature space.

Combining structural relevance and structural redundancy scores in a single score is not straightforward, due to their complimentary semantics. We characterize each feature f in terms of both structural relevance and structural redundancy. To help a domain expert glance the effect a feature may have on their analysis, we combine the two scores in one plot where the x-axis reflects the structural relevance score and the y-axis the structural redundancy. In other words, the x-axis represents the degree to which the reference clustering structure is evident in a single dimension f, while the y-axis shows whether the reference clustering structure can be captured by the rest of the dimensions. These plots can be seen in Figures 4, 5, and 6. They will be explained in detail later in this paper. We present a study using this technique in the following section.

III. Data Analysis in Bioarchaeological Sciences

The presented technique was conceived for an international and interdisciplinary project involving classical archaeology, bioarchaeology, biology, geology, and computer science. The final goal of the project is to generate an isotopic map of the reference region based on isotopic measurements in bone samples of three vertebrate taxa from excavation sites along the Inn-Eisack-Adige passage. A geographic map of the reference region including the sample sites is shown in Figure 1. The envisioned isotopic map will represent the common, local isotopic signatures (or fingerprints) characteristic for a given spatial region. Archaeologists can apply this map to differentiate between local and non-local finds, and to define the place of origin (*provenance analysis* [12], [13]) of the latter in order to answer the aforementioned scientific questions regarding mobility, trade, and cultural transfer. The reason behind this application is that knowledge of the spatial distribution of stable isotopes in the environment allows identifying outliers that represent primarily non-local individuals.

Based on the experience of the domain experts, seven isotopic systems from three elements (oxygen, strontium, and lead) were identified as being potentially relevant for the differentiation between local and non-local finds, and the definition of the place of origin of the latter, i.e., the construction of the envisioned map. From the sites displayed in Figure 1 samples were selected and for each seven isotope ratios were measured. The goal of this study is to identify which of the isotopic systems (oxygen, strontium, lead) to use for provenance analysis in this reference region. The processing of the material and the measurement of isotopic signatures is costly, time consuming, and wastes precious archaeological material. Thus, the design of the underlying data collection (which samples and isotopic systems are measured, etc.) are crucial for the inference of a sound and reliable map.

It should be stressed that the results presented in this study only hold for the specific reference region, i.e., the Inn-Eisack-Adige passage. However, our framework is quite generic and is applicable to other reference regions and/or other isotopic systems and even entirely different data sets from a multitude of disciplines and use cases.

A. A Brief Introduction to Isotopic Mapping

Bioarchaeologists are frequently faced with the task of distilling a relatively simple model from measurements that are tainted by the complexity of the biological processes involved. This is especially characteristic of isotopic mapping. Nevertheless, stable isotopes are indispensable markers for the monitoring of the flow of matter through biogeochemical systems. Isotopes are atoms of the same element that have the same number of protons and electrons, but differ in the number of neutrons. Isotopes are generated, e.g., by the decay of parent isotopes, or by reactions with subatomic particles in the environment. For example, the three stable isotopes of oxygen are ^{16}O, ^{17}O, and ^{18}O. All of these have 8 protons and 8 electrons, but range from 8 (^{16}O) to 10 neutrons (^{18}O). An isotope is called "stable" if it does not decay into another isotope. Oxygen atoms with fewer (e.g. ^{15}O) or more (e.g. ^{19}O) neutrons are unstable and will eventually decay into other stable isotopes.

Differences in the number of neutrons results in different atomic masses and lead to differences in molecular bond strength and vibration energies. This, and the different thermodynamic reactivity of light and heavy isotopes leads to isotopic fractionation (i.e., uneven partitioning of isotopes between source and product). Isotopic fractionation and mixing in an ecosystem generate compartments with characteristic isotopic signatures. For example, evaporation and condensation in the course of hydrological processes lead to predictable distributions of oxygen isotopes in the atmosphere and in precipitation. Isotopic labels, which are shared by certain ecological components such as soil, water, plants, microbia, and animals, have been successfully used for the generation of *isotopic maps* or *isoscapes* for the investigation of landscape ecology. Such isotopic maps representing the common, local isotopic signatures (or fingerprints), can later be applied to distinguish local and non-local finds: a local outlier, i.e., a sample found at location l that has an isotopic fingerprint different to the local fingerprint of l according to the map, is interpreted as non-local. If the isotopic fingerprint of the non-local sample matches the isotopic fingerprint of another location o, it is likely that o is its place of origin. Both the knowledge of outliers and their potential places of origin is very valuable for answering research questions in bioarchaeology. For example, isotopic fingerprints of ivory samples are used to predict the place of orgin of this ivory sample (potentially classifying this sample as illegally harvested) [14].

Isotopic maps are empirically generated by sampling the relevant environmental components and by measuring their isotopic signatures. In bioarchaelogy, such samples are human and animal remains found in archaeological sites. However, due to intricate biological and chemical processes, these samples do not directly reflect the geological characteristics. Examples of such processes include metabolic differences between species and individuals (some of the inter-species differences can be reduced by applying empirically determined formulas), aging, integration over various environmental conditions, weathering of bones, metabolization, etc. Thus the geological characteristics of a region are only one of a few factors contributing to the measured isotope ratios. Since we cannot know the details of the metabolism that crucially influence the isotopic composition of an organism, the only realistic way of modeling the distribution of isotopes in animals found in a region is by building a model based on the measurements associated with them. One way to allow all other influences to average out is to aggregate over samples from spatially close sites. However, the resulting values may not be applicable as a model to a single sample, which is still subject to individual variability as outlined above. In addition, on the one hand, non-local finds at a specific site will most likely impact the aggregated local model in an undesired way. But, on the other hand, the identification of non-local finds requires a reliable local model. As a consequence, the local isotopic fingerprints provided by isotopic maps will never be as reliable as the term fingerprint may suggest, but are always subject to probabilistic interpretation.

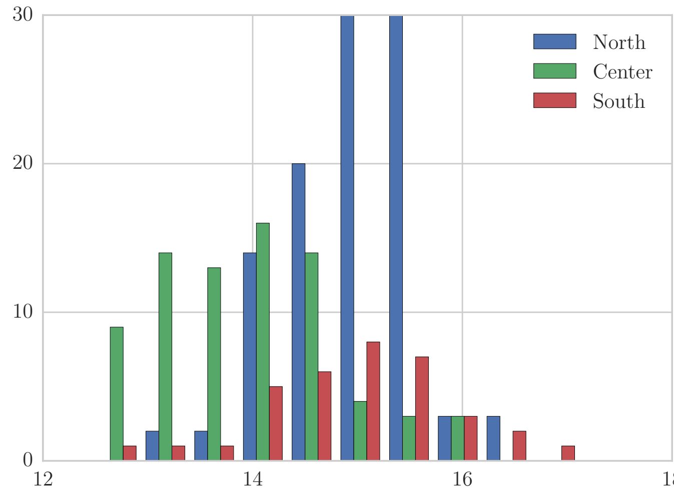


Fig. 3: Distribution of δ^{18}O by region. Although very large regions were picked, there is only a very weak correlation discernable.

Historically, stable isotopes in bioarchaeological finds were measured and simply compared to the known spatial distribution of the isotopic system under study such as ^{87}Sr/^{86}Sr in geological maps, or the climate and habitat dependent distribution of C_3- and C_4-plants which is reflected in the ^{13}C-values of the consumers' tissues. Outliers, detectable by conservative statistics (e.g. [15]), were readily interpreted as immigrant individuals. Very often, this was simply done by measuring one specific isotopic system, e.g. δ^{18}O from phosphate in bones, and manually determining local models and outliers in the univariate plots of the resulting values. However, growing insights into small-scale variabilities in isotopically characterized ecogeographical compartments gave rise to more fruitful discussions on mobility versus migration and trade in the past. Pretty soon it became obvious that measurement of stable isotopes for the reconstruction of migration and trade in bioarchaeology can not be looked at in isolation but rather necessitates collecting a lot of accompanying data (e.g., analysis of not only human, but also animal bones, or soil sampled from the same site) for the assessment of ecogeographical baseline values to account for the small-scale variability in time and space.

One important effect of the complicated nature of this area is that there is no ground truth available. Handling the problem in an unsupervised way means that the investigation cannot rely on an authorized reference, but has to make and use its own references heavily based on assumptions and experiences of the domain experts. This makes this application ideal for a data-driven approach that is underpinned by domain-expertise.

B. The Application: Mobility in the European Alps

The prehistoric transalpine passage following the Inn, Eisack, and Adige rivers from modern-day Germany, through Austria, into Italy is of great archaeological interest for having been an important trade route over the Alps mountains. An international research group has investigated 30 archaeological sites along this route and at its southern and northern ends (see Figure 1) to generate a dataset based on which an isotopic map can be derived for this reference region. The application of this map will leverage the investigation of questions about transfer of humans, goods, and culture through the passage. Of particular interest were animal remains that were uncovered at the examined sites with a focus on archaeological bones of three vertebrate taxa: pig, cattle, and red deer. These three animals have different characteristics in terms of mobility.

In general, it is very time consuming and costly to generate the data an isotopic map is based on. The measurement of an isotopic value requires a complex procedure including, among others, the extraction of a suitable part of the material, cleaning, etc. Typically, the material used for one measurement is destroyed in the analysis and cannot be used for further runs. Thus, one crucial first step towards the establishment of the final map is to decide which isotopic measures to generate. For that purpose, a case study was done with a small set of 217 samples derived from 30 investigated sites. From each investigated specimen, seven isotopes were measured: ^{18}O, ^{86}Sr, ^{87}Sr, ^{204}Pb, ^{206}Pb, ^{207}Pb, and ^{208}Pb. Due to technical particularities of isotope measuring, the strontium (Sr) and lead (Pb) isotopes were measured and recorded as fractions of isotopes of the same element, yielding the fractions $^{87}Sr/^{86}Sr$, $^{208}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, $^{206}Pb/^{204}Pb$, $^{208}Pb/^{207}Pb$, and $^{206}Pb/^{207}Pb$. The oxygen isotope was normalized against ocean water isotope levels and recorded as $\delta^{18}O$. This yields a 7-dimensional feature vector for each recovered sample. In addition to these isotope measurements, each sample was annotated with a spatial description (latitude, longitude, altitude) based on the discovery area. Also, each sample was recognized as one of the three animal species (pig, cattle, and red deer).

Due to its popularity in provenance analysis, we first focused on oxygen as a marker to distinguish between local and non-local finds. A preliminary manual analysis of the δ^{18}O of 118 of the 217 animal bone samples by a domain expert revealed a highly significant correlation ($r = -0.68$) between δ^{18}O and altitude, whereby the averaged δ^{18}O values plotted exactly on the regression between altitude and δ^{18}O in precipitation in the Alps as published by Kern et al. [16]. However, although the δ^{18}O values behaved as expected and, thus, are potentially suitable for provenance analysis as a single marker, in this study it proved impossible to distinguish even rough spatial compartments (such as north, center, and south of the Alps). The histogram of δ^{18}O values shown in Figure 3 illustrates this variability. The samples from the three very coarse compartments north, center, and south derived from a hypothesis of the domain experts are marked in different colors. The same observation was made when considering species-specific fractionation factors. Interindividual variability remained high and did not permit for a firm assignment of individual animals to spatial regions.

These first results already gave the domain experts some very interesting hints and further questions arose. In particular, the domain experts were even more interested in deeper insights into the relevance and redundancy of the measured isotopes. Especially, whether the hypothesis can be confirmed that oxygen isotopes can be omitted or easily replaced by a (set of) other isotopes. Then it would be possible to use datasets of samples where oxygen isotopes could not be measured for the same analyses. These samples are common, because oxygen isotopes are not stable at high temperatures, i.e., in cremated material. Cremated material is very common for human remains in this reference region. The measuring of the other isotopes is slightly easier. If the subset without oxygen is enough for origin prediction, a more detailed model might be derived by augmenting the dataset with human cremation data. Thus, the data science task was to score the measured isotopic features in the dataset of the case study in terms of relevance and redundancy with respect to provenance analysis.

Reference clustering	Description (feature set)
I	all 7 isotopic features ($^{87}Sr/^{86}Sr$,$^{208}Pb/^{204}Pb$,$^{207}Pb/^{204}Pb$,$^{206}Pb/^{204}Pb$, $^{208}Pb/^{207}Pb$,$^{206}Pb/^{207}Pb$, $\delta^{18}O$)
I^{-O}	all isotopic features except oxygen ($^{87}Sr/^{86}Sr$,$^{208}Pb/^{204}Pb$,$^{207}Pb/^{204}Pb$,$^{206}Pb/^{204}Pb$, $^{208}Pb/^{207}Pb$,$^{206}Pb/^{207}Pb$)
S	all 3 spatial attributes ($altitude, latitude, longitude$)
S^{-lon}	all spatial features except longitude ($altitude, latitude$)

TABLE I: Notations for the different subsets of features used to derive reference clusterings.

C. The Results

The defintion of the reference clustering is crucial for our ARI-based feature evaluation presented in the previous section. However, there is no ground truth reference clustering available for the region under inspection. A purely data-driven approach is also not possible since we cannot be sure about the originality of each finding, i.e., we do not know if a bone found a specific site s in fact originates from s or is a non-local outlier. As a consequence, even if we explore local isotopic outliers within each site, we are not sure if the outliers are non-local finds or local ones. However, the domain experts have some assumptions and hypotheses available about possible spatial compartments that could be used to derive potentially plausible approximations of the ground truth. Thus, we follow a mixture between a data driven approach enriched by domain expertise.

1) Reference clusterings: Instead of using just one potential reference clustering, we investigated several possible definitions for the reference clustering based on the available features, in close collaboration with the domain experts. In other words, we generated reference clusterings using a data driven approach based on clustering, but rely on the domain experts for deciding which features were used for the clustering. As a result of this process we decided on multiple feature spaces to generate reference clusterings: from containing all isotope and spatial features to containing only single domain features, i.e., isotopes or spatial coordinates. In the following, the set $I := \{^{87}Sr/^{86}Sr, ^{208}Pb/^{204}Pb, ^{207}Pb/^{204}Pb, ^{206}Pb/^{204}Pb, ^{208}Pb/^{207}Pb, ^{206}Pb/^{207}Pb, \delta^{18}O\}$ denotes all isotopic features, $I^{-O} := \{^{87}Sr/^{86}Sr, ^{208}Pb/^{204}Pb, ^{207}Pb/^{204}Pb, ^{206}Pb/^{204}Pb, ^{208}Pb/^{207}Pb, ^{206}Pb/^{207}Pb\}$ denotes all isotopic features except oxygen. In addition, we use $S := \{altitude, latitude, longitude\}$ for all spatial attributes and $S^{-lon} := \{altitude, latitude\}$ refers to the spatial attributes without longitude (see Table I for an overview). We list the different set-ups for the feature spaces that we used to generate the reference clusterings based on the input of the domain experts below. For this first set of experiments, the set of features under investigation is always the set of isotopic features, i.e., $F_v := I$.

$F_0 = I \cup S$ **(Isotopes + Spatial)** The feature space consists of all available isotopic features and spatial features. This is the most information we have and, thus, serves as the starting point of the study.

$F_0 = I \cup S^{-lon}$ **(Isotopes + (latitude, altitude))** From the spatial attributes only those that have been found to have an effect on the isotopes are retained, namely altitude and latitude. Since the passage under inspection is mostly north/south, the domain experts expect that longitude has only minor influence on the spatial compartments.

$F_0 = I$ **(Isotopes only)** The feature space consists only of the isotopic features. There is no spatial influence. Such a feature space is typically used for fingerprinting and predicting the origin of new samples (with unknown spatial coordinates).

In a second analogously conducted series of experiments, oxygen was removed from the reference clusterings since the domain experts wanted to test the hypothesis that oxygen is much less relevant than other isotopes in this reference region and for this sample selection. This has been observed in other provenance studies. Especially the sample selection using a mix of three different species may have a blurring impact on the $\delta^{18}O$-values according to the domain experts. Analogously, the set of features under investigation is always the set of isotopic features without oxygen, i.e., $F_v = I^{-O}$. The resulting configurations are similar to the four alternatives listed above:

$F_0 = I^{-O}S$ **(Isotopes (except oxygen) + Spatial)** The feature space consists of all isotopes minus oxygen and all spatial features.

$F_0 = I^{-O}$ **(Isotopes only (except oxygen))** Only the isotope description, without the oxygen feature.

$F_0 = I^{-O} \cup S^{-lon}$ **(Isotopes (except oxygen) + (latitude + altitude))** Isotope description, without the oxygen feature and spatial coordinates except longitude.

A supplementary feature space to be used as a reference clustering is purely spatial:

$F_0 = S$ **(Spatial only)** The feature space consists only of spatial coordinates. Isotopic values do not play any role and findings from spatially close sites are considered to be the same cluster (compartment). This ground truth scenario must be complemented by a corresponding set of investigated features, i.e., $F_v = I$, and $F_v = I^{-O}$.

2) Experiments: For each of the feature spaces described above, we apply EM to derive the reference clustering and we evaluate how each isotope attribute "contributes" to the corresponding reference clustering. We illustrate these results in the structural relevance-vs-structural redundancy plots presented in the previous section. For the EM, the number of clusters was selected by cross-validation as in the Weka data mining framework [17]. When examining the presented reference attribute sets, we chose $F_v = I$ or $F_v = I^{-O}$ to reflect the isotopes in the reference attributes. That is, where F_0 contains I, F_v becomes I, where F_0 contains only I^{-O}, F_v becomes I^{-O}. A special case is $F_0 = S$, which does not contain any isotopes to compare with. In these scenarios, we investigated both $F_v = I$ and $F_v = I^{-O}$ for completeness.

The results of reference clusterings containing isotopes including oxygen are presented in Figure 4, experiments with isotopes excluding oxygen are presented in Figure 5, and those with only spatial attributes are presented in Figure 6. Regarding the ARI values, a score of zero indicates random behavior while a score of one indicates identical clusterings.

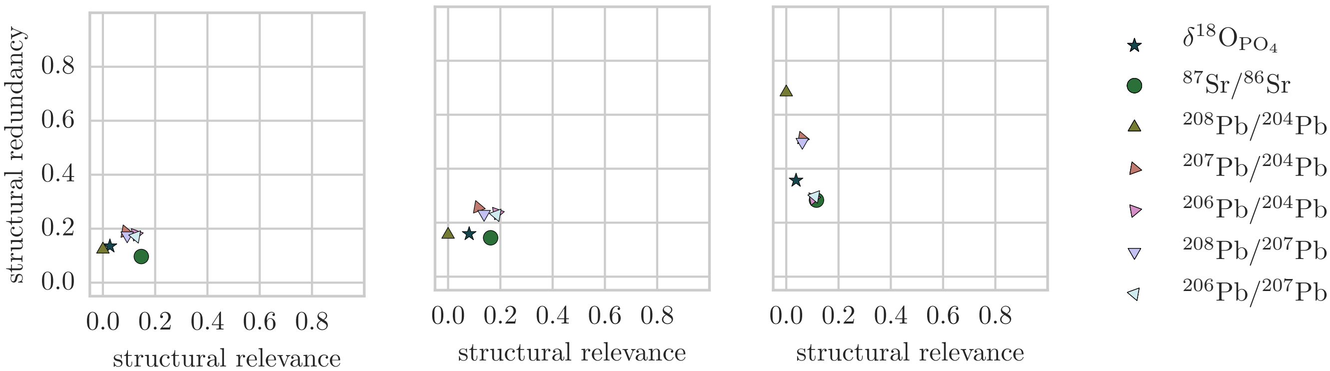
In the following we discuss the individual experiments showing structural redundancy and structural relevance for all described reference feature sets and discussing potential explanations for the observed values.

a) $F_0 = I \cup S$, $F_v = I$: See Figure 4a. *Strontium* is the most prominent attribute as it has the highest structural relevance score and the lowest structural redundancy score. *Lead* isotopes depict a similar behavior, scoring average relevance and redundancy scores. An exception is $^{208}\mathrm{Pb}/^{204}\mathrm{Pb}$, which has a very low relevance; a closer inspection of the results shows that a clustering based on $^{208}\mathrm{Pb}/^{204}\mathrm{Pb}$ only places all instances in the same cluster, i.e., the values in this feature follow one Gaussian distribution. *Oxygen* has also a very low relevance score.

b) $F_0 = I \cup S^{-lon}$, $F_v = I$: See Figure 4b. F_0 now includes no longitude information. There is little difference in the rankings compared to the IS case, although the scores are higher. An interesting change is the repositioning of oxygen: its redundancy became lower and relevance became higher.

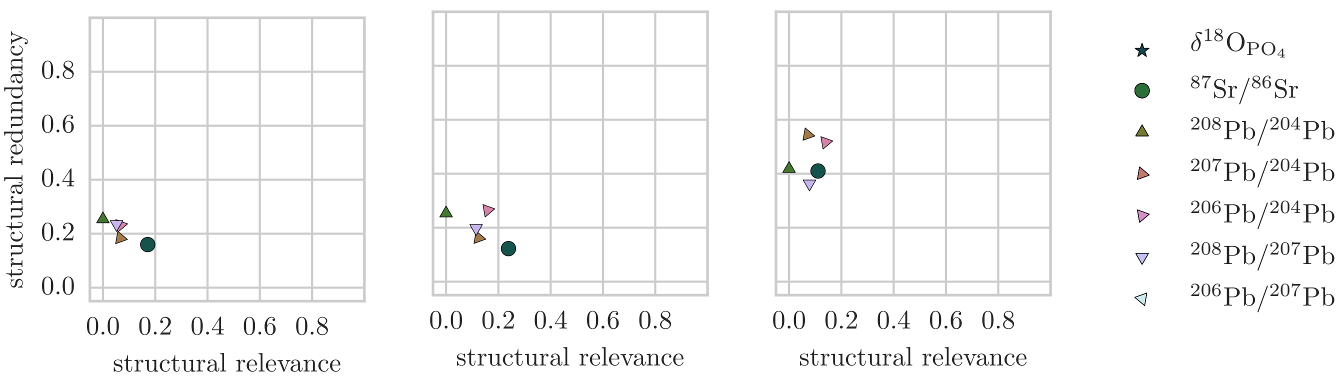
c) $F_0 = I$, $F_v = I$: See Figure 4c. The removal of all spatial information from F_0 pits the isotopes against each other. This might be due to the better quality clusterings we obtain by also employing spatial information. Strontium is still the top relevant isotope, however two lead isotopes score very close, namely $^{206}\mathrm{Pb}/^{207\mathrm{Pb}}$ and $^{206}\mathrm{Pb}/^{204}\mathrm{Pb}$. Both isotopes re-positioned in the plot after the removal of the spatial information from the reference clustering. In particular they became more relevant and less redundant. Also, the redundancy of oxygen increased.

d) $F_0 = S$, $F_v = I$: See Figure 6a. This scenario tests how well the isotope's structure lines up with the spatial structure. We expect very little alignment as the spatial structure



(a) $F_0 = I \cup S, F_v = I$ (b) $F_0 = I \cup S^{-lon}, F_v = I$ (c) $F_0 = I, F_v = I$

Fig. 4: Structural relevance-vs-structural redundancy plots using reference clusterings with all isotope features.



(a) $F_0 = I^{-O} \cup S, F_v = I^{-O}$ (b) $F_0 = I^{-O} \cup S^{-lon}, F_v = I^{-O}$ (c) $F_0 = I^{-O}, F_v = I^{-O}$

Fig. 5: Structural relevance-vs-structural redundancy plots using reference clusterings with all isotopes except oxygen.

will be dominated by the density of sample sites, which the isotope values reflect indirectly at best. The lead isotopes have very low redundancy and relevance scores, indicating that they neither reflect the spatial structure, nor does their complimentary feature space do so. Strontium seems to reflect all the structure: the sets of isotopes that contains strontium achieve a median score of 0.08 and strontium by itself achieves a score of 0.13. All other isotopes have relevance scores around zero (oxygen scoring highest at 0.04).

e) $F_0 = I^{-O} \cup S$, $F_v = I^{-O}$: See Figure 5a. Without oxygen, strontium is again the most prominent attribute, whereas the relevance of lead decreases.

f) $F_0 = I^{-O} \cup S^{-lon}$, $F_v = I^{-O}$: See Figure 5b. Compared with the previous scenario containing the entire set of spatial attributes, strontium retains very similar scores, but some lead isotopes' relevance increases. This indicates a stronger role of those lead isotopes in the formation of the structure, possibly because longitude supports other structural elements that are now being expressed less strongly.

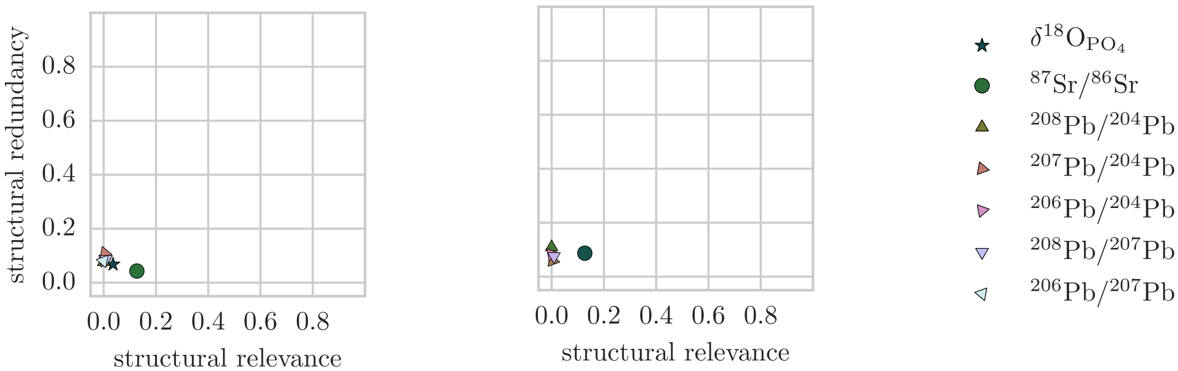
g) $F_0 = I^{-O}$, $F_v = I^{-O}$: See Figure 5c. Removal of all spatial information (and oxygen), affects the ranking of strontium. Lead isotopes are now more relevant comparing to strontium, but still more redundant than strontium.

h) $F_0 = S$, $F_v = I^{-O}$: See Figure 6b. If oxygen is omitted, the situation changes only marginally compared to the original setup with $F_v = I$. This indicates that oxygen had little influence on the structure of the isotope space, consistent with the analysis above.

3) Discussion: We already pointed out that each experiment refers to a specific reference clustering and therefore it is not straightforward to compare scores between them. However we can draw some conclusions about the dataset from the interpretation of all experiments.

First of all, it is clear that the choice of reference clustering influences the ranking of different isotopes with respect to their structural relevance and structural redundancy. There are however some features which are repeated across different configurations. In particular, there is a much better separation in the rankings when spatial coordinates are considered, cf. Figure 4a and 4b. A possible explanation is that the reference clustering is much more differentiated when considering the complete feature space. A similar observation holds when we remove oxygen from the feature space, cf. Figure 5. The worst distinction is manifested when we consider all isotopes, including oxygen, cf. Figure 4c.

Regarding the behavior of the different isotopes, *strontium* and *lead* are the top *structurally relevant* isotopes, i.e., they



(a) $F_0 = S, F_v = I$

(b) $F_0 = S, F_v = I^{-O}$

Fig. 6: Structural relevance-vs-structural redundancy plot using reference clustering on spatial data. The investigated clustering are based on all isotopes (a) and all isotopes except oxygen (b).

display higher values in the structural relevance axis. This implies that in isolation these isotopes manage to capture most of the reference clustering structure. *Oxygen* depicts a low structural relevance score, meaning that alone it is not a good indicator of the reference clustering.

With respect to *structural redundancy*, strontium has the lowest redundancy, implying that the information in strontium is not replicated by some other isotope or combination of isotopes in the dataset. The lead isotopes display high redundancy as expected since we have five different lead isotopes in our dataset. Related behavior within the lead group is expected, because the lead isotopes used in this study are measured relatively against baseline isotopes ^{204}Pb and ^{207}Pb. This implies that each isotope is measured multiple times as part of different fractions. Although they are measured independently, these fractions can be mathematically converted into one another by multiplying two of each data point's measurements and the derived values are generally very close. Accordingly, lead isotope ratios generally score higher on the redundancy score than other isotopes. This multiplicative relationship is not clearly reflected in the data structure and the redundancy scores of lead isotopes are therefore not exceptionally high. It is interesting to note that the lead isotope ratios generally change uniformly and that correlations between the scores of fractions that share an isotope can only be observed in a few cases. Two that do behave similarly are ^{206}Pb/^{207}Pb and ^{206}Pb/^{204}Pb. There seems to be no obvious explanation for why these systems in particular show this behavior, but it may indicate that stable lead isotopes can be used for provenance analysis. This warrants further research.

Overall low relevance scores indicate that no isotope alone reflects the full structure of the data. This supports the emerging trend to use multivariate analyses in the domain sciences.

The bad scores achieved by all isotopes against the reference clustering including only spatial coordinates illustrates that there is no trivial correspondence between the two domains, isotope and spatial. Domain knowledge suggests a connection, but it is not pronounced enough to be automatically reflected by the isotope feature set. Therefore the combination of both domains to extract a spatially coherent isotope map is also not trivial and will require more complex models.

D. *Insights*

Our study resulted in two major insights that were previously uncertain and represented major added values for the domain experts.

Insight 1: *A multivariate isotopic fingerprint is needed instead of a univariate analysis relying on oxygen only.*
Our analysis showed that despite its popularity, oxygen does not provide exceptional structure to the dataset (average structural relevance), nor is it unique in the role it plays (no exceptionally low structural redundancy values). Thus, at least in this reference region, provenance studies based solely on oxygen is bound to fail. On the other hand, the implication from our results is that the envisioned isotopic map can benefit strongly from a multi-isotopic fingerprint that includes strontium and lead isotopes as well.

Insight 2: *Omission of oxygen in the isotopic fingerprint does not considerably decrease the quality of the fingerprinting.*
Oxygen did not show a particularly low redundancy. Its redudancy scores were always comparable with other isotopes, reaching values of up to 35%. This indicates that oxygen does not play an exceptional role in the data's structure and that other isotopes can provide much the same information as oxygen. Its low relevance score indicates that oxygen does not dominate the structure (i.e., other isotopes are needed).

Implications and Added Values: The fact that oxygen seems not very relevant to provenance analysis in the reference region, opens up several opportunities.

- Although the inclusion of oxygen does not seem to diminish the clustering results, its omission also has little negative impact. This makes oxygen a potential candidate to save costs and time when generating the data source on which an isotopic map is based.
- So far, the isotopic map was designed to rely on animal bones only. Including human remains would be generally

beneficial but ancient human remains in this reference region typically are cremated, and, thus oxygen values cannot be derived. The low relevance of oxygen opens up the possibility to explore this cremated material on a larger scale.

IV. Conclusion

This paper presented a technique for domain scientists to assess the relevance of features for analysis. The technique's purpose is to inform decisions about features, such as whether to record a variable in the first place, as well as guide further investigations into the role of a feature. After analysis, domain scientists are presented with two scores for each isotope: the structural relevance, which indicates to what degree the data's structure is represented in a given feature, and the structural redundancy, which indicates how much of the data structure is lost without the feature.

By splitting the result into two independent scores (structural relevance and structural redundancy) we allow domain scientists to grasp two important orthogonal properties of the data that could otherwise not be discerned from univariate and bivariate visualizations. A variable that is structurally relevant, but redundant, may still be less important than one that is structurally less relevant, but cannot be replaced by a combination of different isotopes, or the other way around. In low-dimensional datasets individual variables are expected to be generally more relevant than in higher-dimensional ones. However, no single variable is indispensible if multi-variate analysis is employed. Indeed if the analysis could be based on only a single variable, multi-variate analysis would not be necessary for the application at hand.

In an application context these measurements inform further investigations of the role of features in domain models. In the presented case study, domain scientists were presented with scatter plots of the structural relevance and structural redundancy scores of each isotope system in an archaeological dataset. The presented study was only an early step towards the overall goal of the interdisciplinary research project of mobility and cultural transfer in the past in a specific reference region. This analysis was important since it revealed many new insights for the domain experts, mainly in terms of how the data should be generated (e.g., which findings to include into the analysis, which isotopic values should be measured or can be omitted, etc.) so that a reliable map can be derived.

Analysing the presented data to generate the aspired isotopic map presents further data science challenges. Suitable methods will be needed to identify the small scale compartments with characteristic isotopic fingerprints. Based on these compartments and their characteristic fingerprints, a predictive model needs to be derived in order to classify local findings and non-local findings as well as the places of origin for the latter. To be useful to domain scientists, a visualization of the isotopic map, i.e., of a multi-dimensional fingerprint, is planned that allows for some insights without the full complexity of the underlying model. Finally, the resulting isotopic map must be extendable to surrounding regions like other alpine passages in Austria, Switzerland, and France as well to other archaeological strata.

In this study, we have demonstrated that datascience techniques like the one presented in this paper can generate new insights, inputs, and impulses for domain sciences.

Acknowledgement

This work is supported by the German Research Foundation within the interdisciplinary DFG Research Group FOR 1670 "Transalpine mobility and cultural transfer". Details at: http://www.en.for1670-transalpine.uni-muenchen.de

References

[1] G. Bowen, "Isoscapes: Spatial pattern in isotopic biogeochemistry," *Annual reviews of earth and planetary science*, pp. 161–187, 2010.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[3] F. Keinosuke, "Introduction to statistical pattern recognition," *Academic Press Inc*, 1990.

[4] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997, pp. 532–539.

[5] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[6] M. Mauder, E. Ntoutsi, P. Kröger, and G. Grupe, "Data mining for isotopic mapping of bioarchaeological finds in a central european alpine passage," in *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*. ACM, 2015, p. 34.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[8] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[9] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[10] G. W. Milligan and M. C. Cooper, "Methodology review: Clustering methods," *Applied psychological measurement*, vol. 11, no. 4, pp. 329–354, 1987.

[11] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[12] J. Ericson, "Strontium isotope characterization in the study of prehistoric human ecology," *Journal of Human Evolution 14*, pp. 503–14, 1985.

[13] T. Molleson, D. Eldridge, and N. Gale, "Identification of lead sources by stable isotope ratios in bones and lead from Poundbury Camp, Dorset," *Oxford Journal of Archaeology 5*, pp. 249–253, 1986.

[14] S. Ziegler and D. Jacob, "Development of a spatial reference database for ivory," *TRAFFIC Bulletin*, vol. 23, no. 1, 2010.

[15] G. Grupe, T. Price, P. Schröter, F. Söllner, C. Johnson, and B. Beard, "Mobility of Bell Beaker people revealed by strontium isotope ratios of tooth and bone: a study of southern Bavarian skeletal remains," *Applied Geochemistry 12*, pp. 517–525, 1997.

[16] Z. Kern, B. Kohán, and M. Leuenberger, "Precipitation isoscape of high reliefs: interpolation scheme designed and tested for monthly resolved precipitation oxygen isotope records of an Alpine domain," *Atmospheric Chemistry and Physics 14*, pp. 1897–1907, 2014.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "EM," http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html, 03 2015, [Online; accessed 2015-03-10].