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

Markus Mauder[1], Eirini Ntoutsi[2], Peer Kröger[1], Christoph Mayr[3], Gisela Grupe[4], Anita Toncala[4], and Stefan Hölzl[5]

[1]Institute for Informatics, Data Science Lab, Ludwig-Maximilians-Universität München, Germany

[2]Faculty of Electrical Engineering and Computer Science, Leibniz Universität Hannover, Germany

[3]Institute for Geography, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

[4]Bio-Center, Ludwig-Maximilians-Universität München, Germany

[5]RiesKraterMuseum Nördlingen, Germany

12th International Conference on eScience
2016-10-25

Transalpine mobility and cultural transfer

Research Unit of the German Research Foundation (FOR 1670)

Establishment of an isotopic fingerprint for bioarchaeological finds, especially cremations, and its application to archaeological and cultural-historical problems of the Late Bronze Age until Roman Times

An interdisciplinary project of the ArchaeoBioCenter LMU

ARCHAEOBIOCENTER

BAYERISCHES LANDESAMT FÜR DENKMALPFLEGE

archäologische staatssammlung münchen

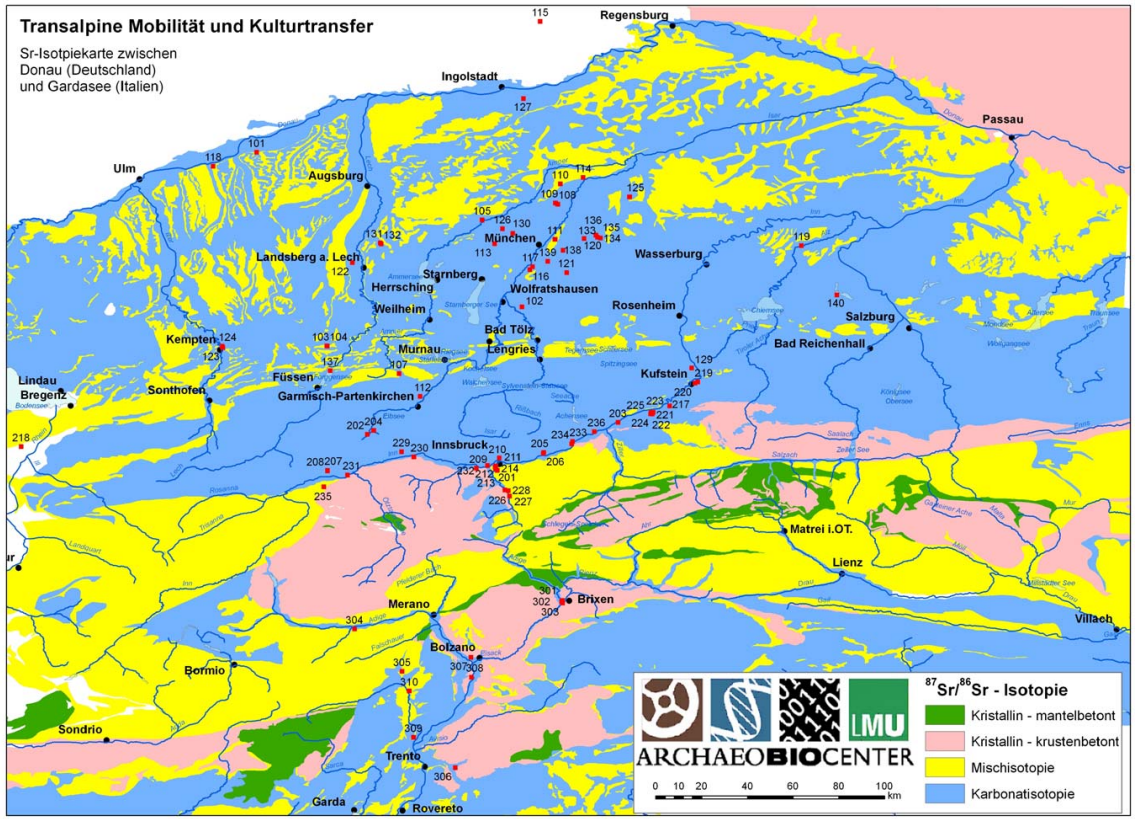
staatliche naturwissenschaftliche sammlungen bayerns

Research institutions beyond university walls

FOR 1670

Project goal: *isotopic fingerprint for bioarchaeological finds*

- build a model that explains and predicts the spatial distribution of this data ("fingerprint")
- using *stable isotope data* from bioarchaeological finds



Data

What is "stable isotope data"?

isotope a "flavor" of an element (different number of neutrons)

stable does not spontaneously change "flavor"

Data

Remains of humans and animals (three species) were analyzed.
The following isotope ratios were measured:

- $^{208}\mathrm{Pb}/^{204}\mathrm{Pb}$
- $^{207}\mathrm{Pb}/^{204}\mathrm{Pb}$
- $^{206}\mathrm{Pb}/^{204}\mathrm{Pb}$
- $^{208}\mathrm{Pb}/^{207}\mathrm{Pb}$
- $^{206}\mathrm{Pb}/^{207}\mathrm{Pb}$
- $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$
- $^{18}\mathrm{O}/^{16}\mathrm{O}$

Oxygen

Oxygen isotopes can change under the influence of high temperatures.
But (from the project description):

> *[Analyze] bioarchaeological finds*, **especially cremations**, . . .

$\rightarrow$ no usable oxygen measurements for human data (which is about half the data set)

Questions from Domain Scientists

Domain scientists have been discussing the following questions:

- What is the role of oxygen in the model of the sample distribution?
- Can we omit oxygen from the analysis and combine the datasets?

Questions from Domain Scientists

Domain scientists have been discussing the following questions:

- What is the role of oxygen in the model of the sample distribution?
- Can we omit oxygen from the analysis and combine the datasets?

Many more questions about the attributes:

- If we want to include spatial data (build a map), how is the distribution affected?
- Which isotopes can be left out until the model becomes different? e.g. is there any value in including all Pb isotopes?

$\rightarrow$ find a way to compare different isotope feature sets' ability to be used as fingerprint

Idea

Compare the effect of modeling the data based on different attribute subsets.

Steps

1. Make a model using the reference attribute set
2. Make a model using the evaluation attribute set
3. Compare the effect of the model

$\rightarrow$ What is an appropriate model?

Target model

Geologists: isotope distributions follow Gaussian models
→ train a Gaussian Mixture Model that explains the data (and makes sense spatially)

EM algorithm

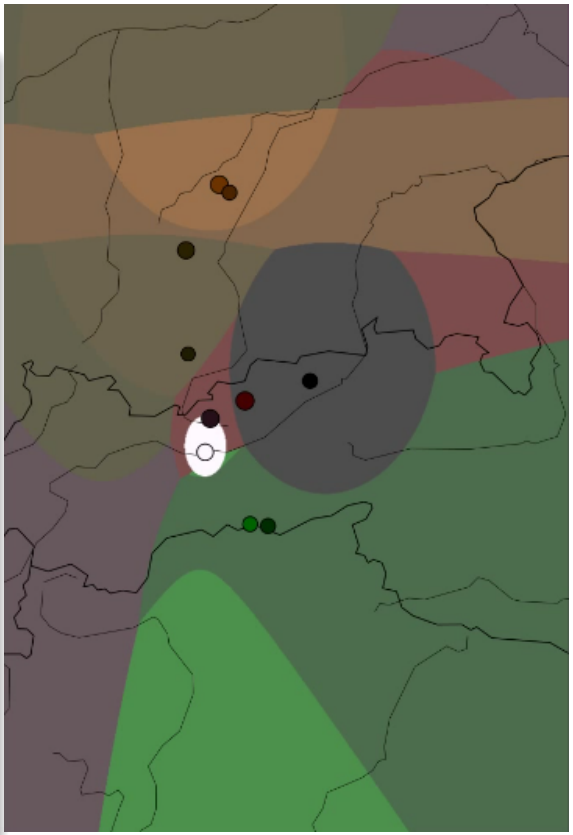
input samples, number of clusters k

initialize build initial GMM (k models)

repeat

1. assign probabilities to (sample, cluster)-tuples based on GMM
2. update the current GMM from the current probabilities

output GMM and probability of assignment of each sample to each cluster



→ Compare the results

Adjusted Rand Index

Goal: Compare the cluster assignments.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where

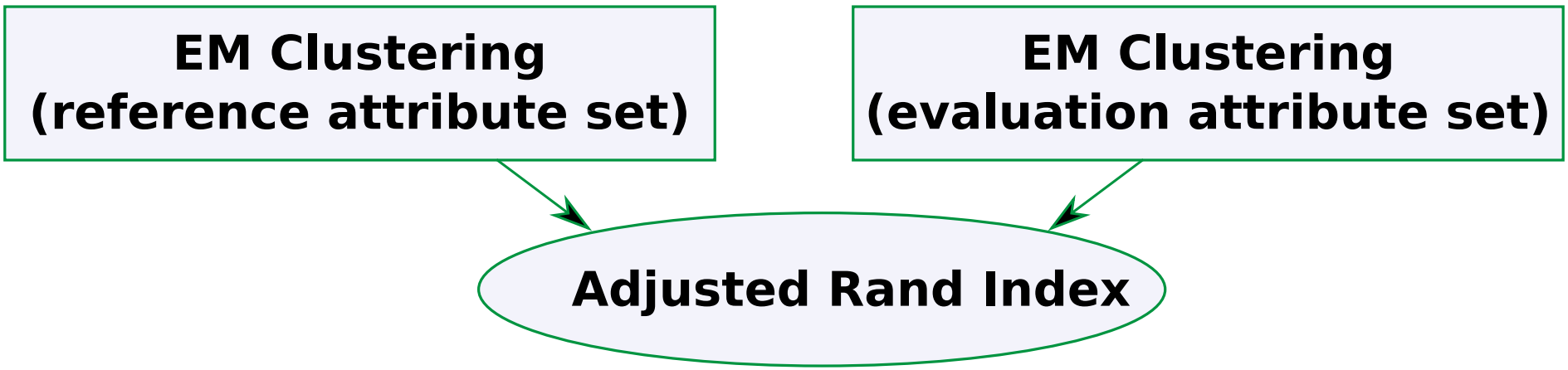
- n_{ij} is the number of points that are in cluster i in clustering 1 and in cluster j in clustering 2,
- a_i is the number of points in cluster i in clustering 1, and
- b_i is the number of points in cluster i in clustering 2.

Summary: comparing attribute sets

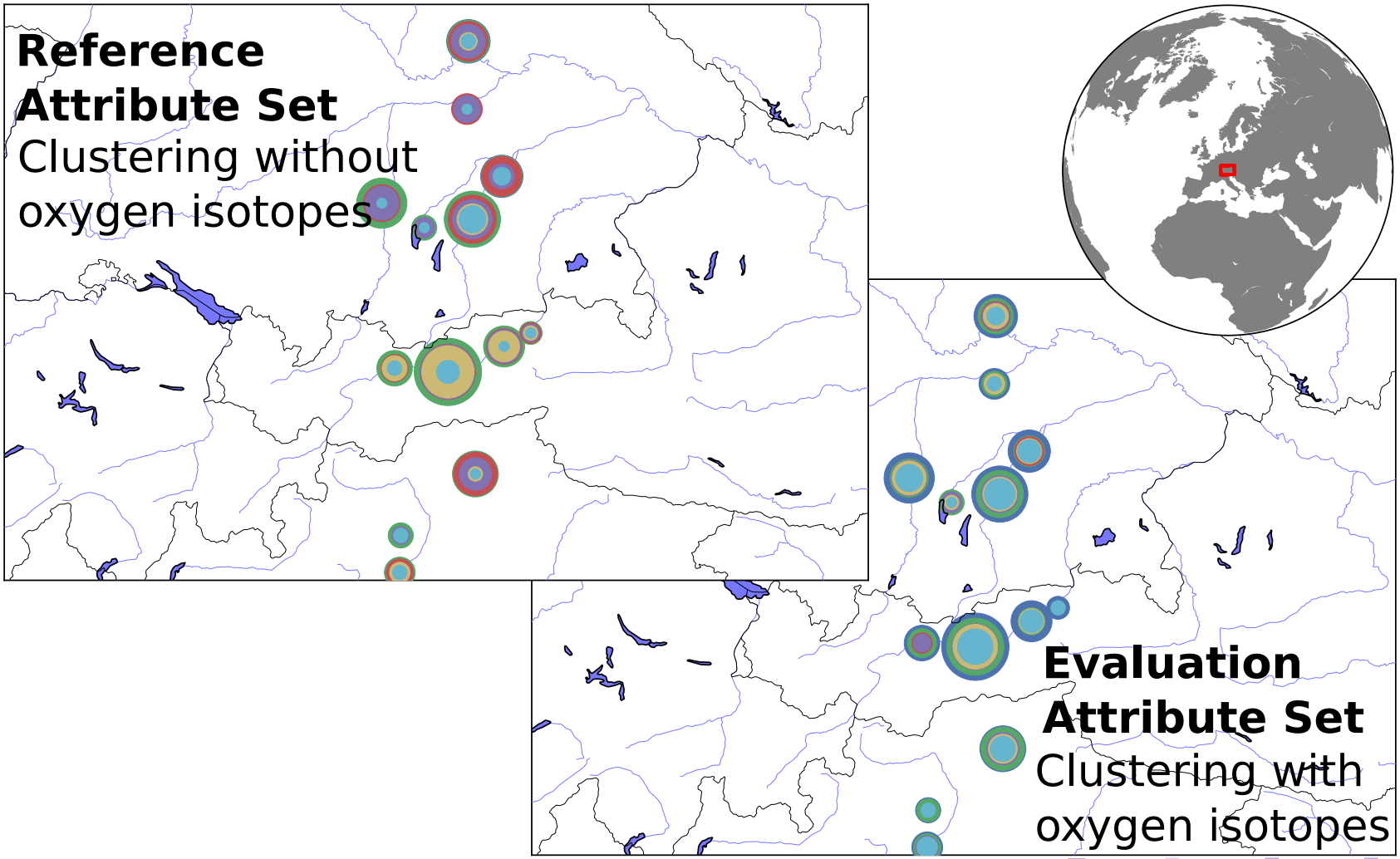
input reference attribute set

input evaluation attribute set

output similarity of result model



Example: ML cluster assignment based on GMM of different attribute sets



Translating domain scientists' questions

Rephrase domain scientists' questions as questions about the differences between attribute sets.

For a single attributes (oxygen):

- clustering based on the single isotope, vs
- clustering based on all but the one attribute

Different reference attribute sets:

- how similar are results with/without spatial information?
- how similar are results with/without different isotope subsets?

Application to domain scientists' questions

Let's try and figure out the answer to the original questions:

- What is the role of oxygen in the model of the sample distribution?
- Can we omit oxygen from the analysis and combine the datasets?

For different reference attribute sets A, test the influence of each isotope $a \in A$ by:

- basing the clustering on a alone (*structural relevance*)
- basing the clustering on $A \setminus \{a\}$ (*structural redundancy*)

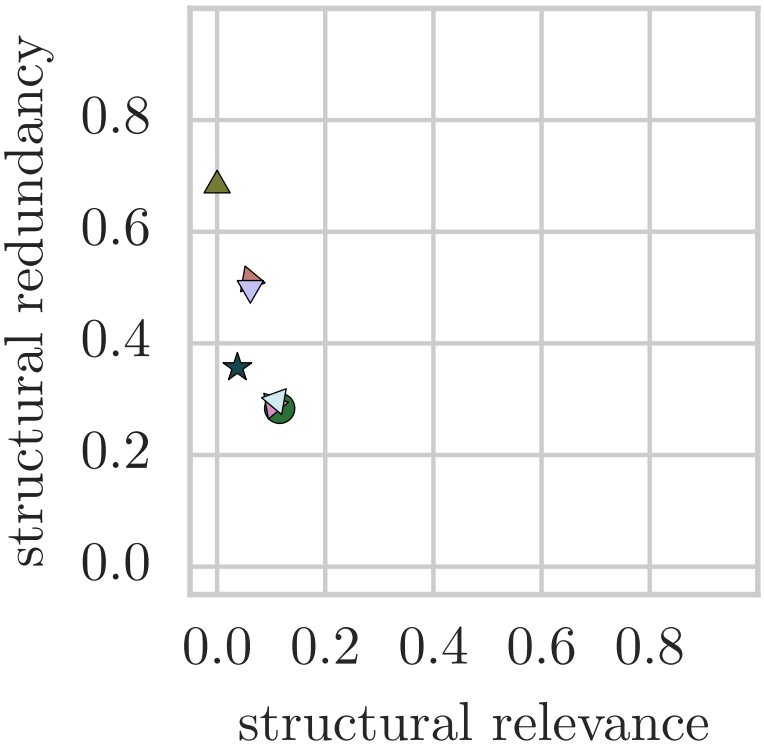
Available attributes to test different scenarios:

I isotope ratios

S spatial information $\{lat, lon\}$

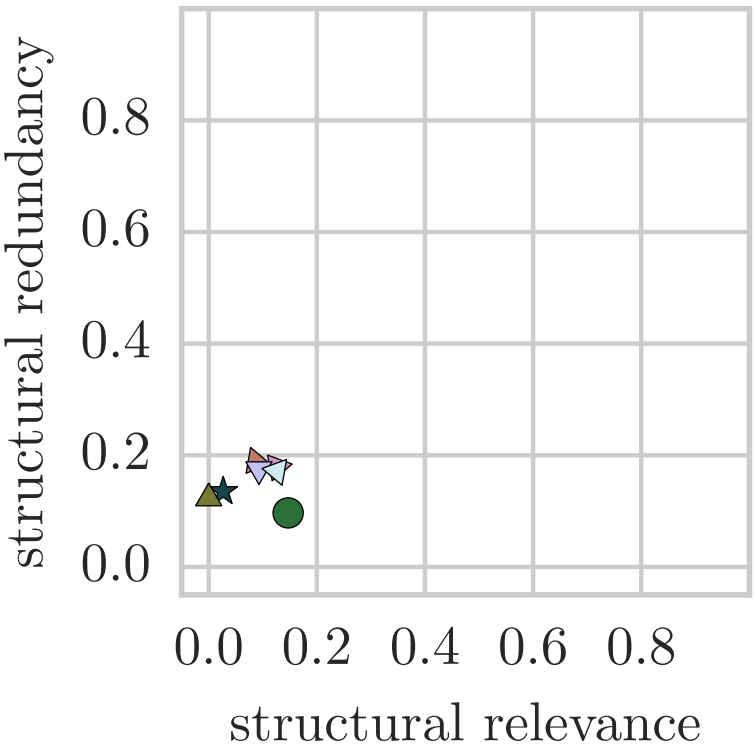
Example: *I*

Same evaluation and reference attribute sets: the set of all isotopes *I*.



Example: *IS*

Reference attribute set is the set of all isotopes and spatial data $I \cup S$.
Evaluation attribute set is the set of all isotopes I.



Summary

- Archaeology is being eScience'd
- The presented project investigates the place of origin of animals and humans.
- This study was concerned with the role of individual attributes in the modeling of isotope distributions
- (Bio-)archaeologists: rather have a larger dataset than oxygen

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

Markus Mauder[1], Eirini Ntoutsi[2], Peer Kröger[1], Christoph Mayr[3], Gisela Grupe[4], Anita Toncala[4], and Stefan Hölzl[5]

[1]Institute for Informatics, Data Science Lab, Ludwig-Maximilians-Universität München, Germany

[2]Faculty of Electrical Engineering and Computer Science, Leibniz Universität Hannover, Germany

[3]Institute for Geography, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

[4]Bio-Center, Ludwig-Maximilians-Universität München, Germany

[5]RiesKraterMuseum Nördlingen, Germany

12th International Conference on eScience
2016-10-25