Empirically exploring the effect of oxygen on the isotopic mapping of cremated and uncremated bones of a Central European Alpine passage.

Markus Mauder, Eirini Ntoutsi, Gisela Grupe, Peer Kroeger

6th International Symposium on Biomolecular Archaeology

27. - 30.8.2014, Basel, Switcherland

1 LMU Institute for Informatics,
Ludwig-Maximilians-Universität (LMU)
München, Germany

2 ARCHAEOBIOCENTER

Outline

- A short overview of our project and the important of isotopic fingerprinting
- Employing data mining for isoscaping
- Studying the effect of oxygen on the extracted models
- Discussion and outlook

Transalpine mobility and cultural transfer project

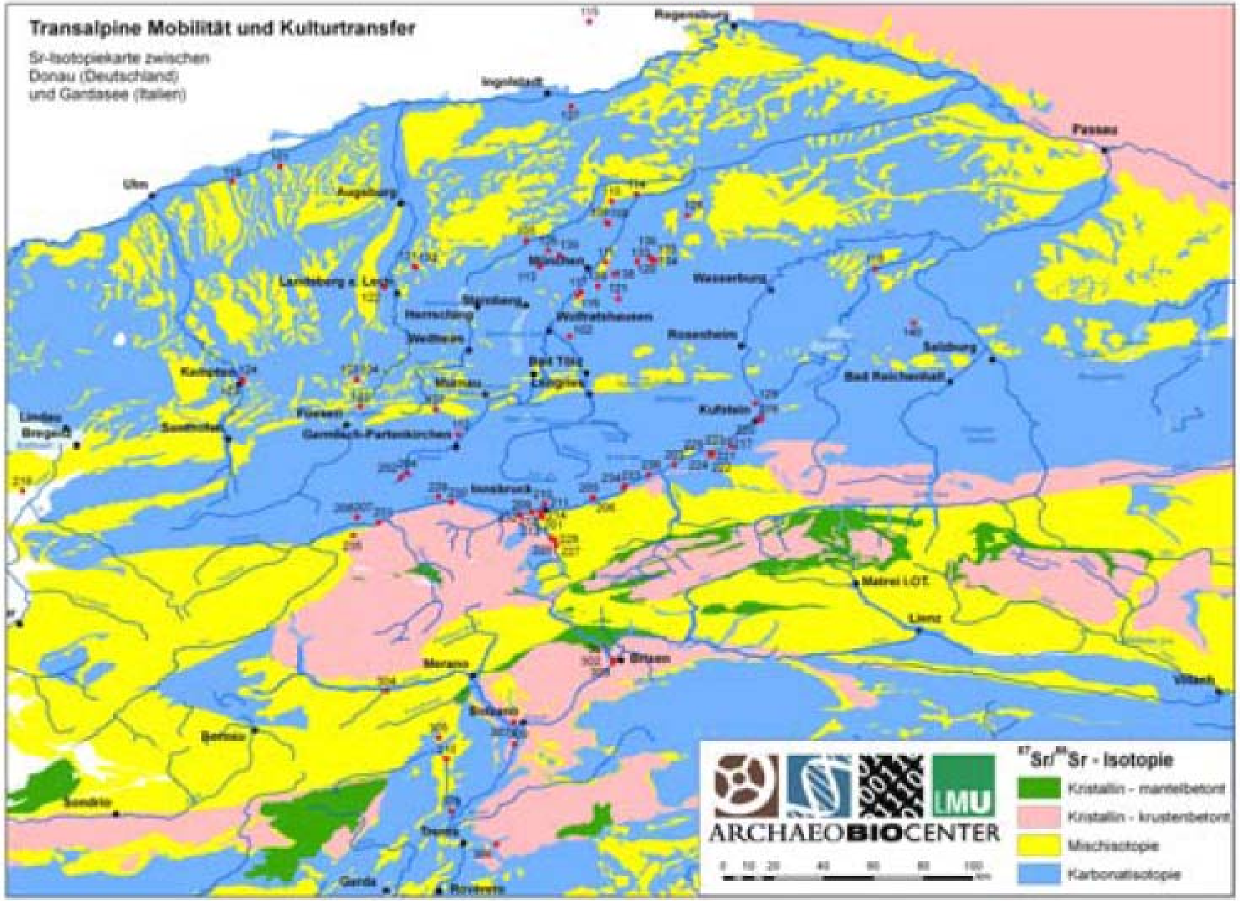
- An interdisciplinary project of the Archaebiocenter, LMU, Munich
- Research Unit of the German Science Foundation, DFG (FOR 1670)

- Goal: Establishment of an isotopic fingerprint for bioarchaeological finds, especially cremations, and its application to archaeological and cultural-historical problems.

- Reference region: the transalpine Inn-Eisack-Etsch-Brenner passage. Specific archaeological contexts from Late Bronze Age until Roman times.

- Project www: *http:/www.for1670-transalpine.uni-muenchen.de*

Isotopic mapping



- Samples: animal findings
- Isotopes considered: Strondium, Lead, Oxygen

Building an isotopic fingerprint

- Isoscaping is a task of paramount importance in order to
  - describe/ “understand” an area
  - predict the most probable (spatial) origin of new samples

- Two data mining approaches towards this goal:

  1) The supervised way:

  Given the locality of the samples, can we generate a model that captures the key characteristics of the localities and is able to predict the locality of new samples?

  - Spatial coordinates of the samples are also part of the model.
  - The list of localities (problem classes) is predefined.

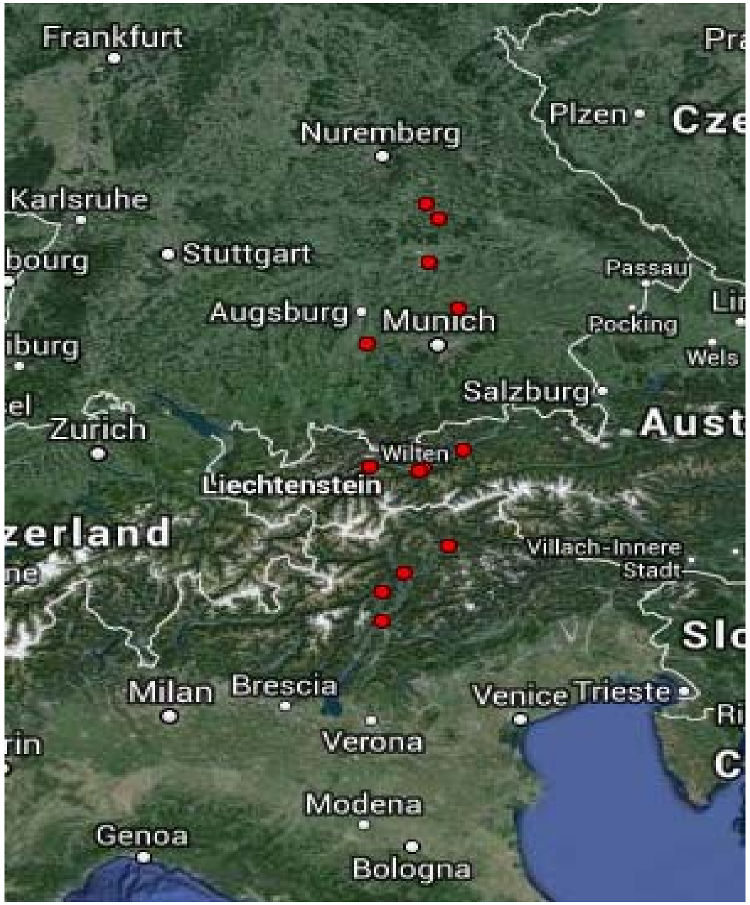
  2) The unsupervised way

  Can we group samples based solely on their isotopic values and check how the extracted *isotopic-clusters* are *spatially scattered* ?

  - Only isotopic values of the samples are used for clustering.
  - Their coordinates are used for spatial validation/ exploration.

Our data

- Dataset consists of ~100 samples
- Each sample described in terms of:
  - Spatial coordinates (lat, long)
  - 3 isotopes (Sr, Pb, O) and
  - 7 isotope ratios
    - 87Sr/86Sr
    - 208Pb/204Pb
    - 207Pb/204Pb
    - 206Pb/204Pb
    - 208Pb/207Pb
    - 206Pb 207Pb
    - 18OPO4



Geographic distribution of the samples

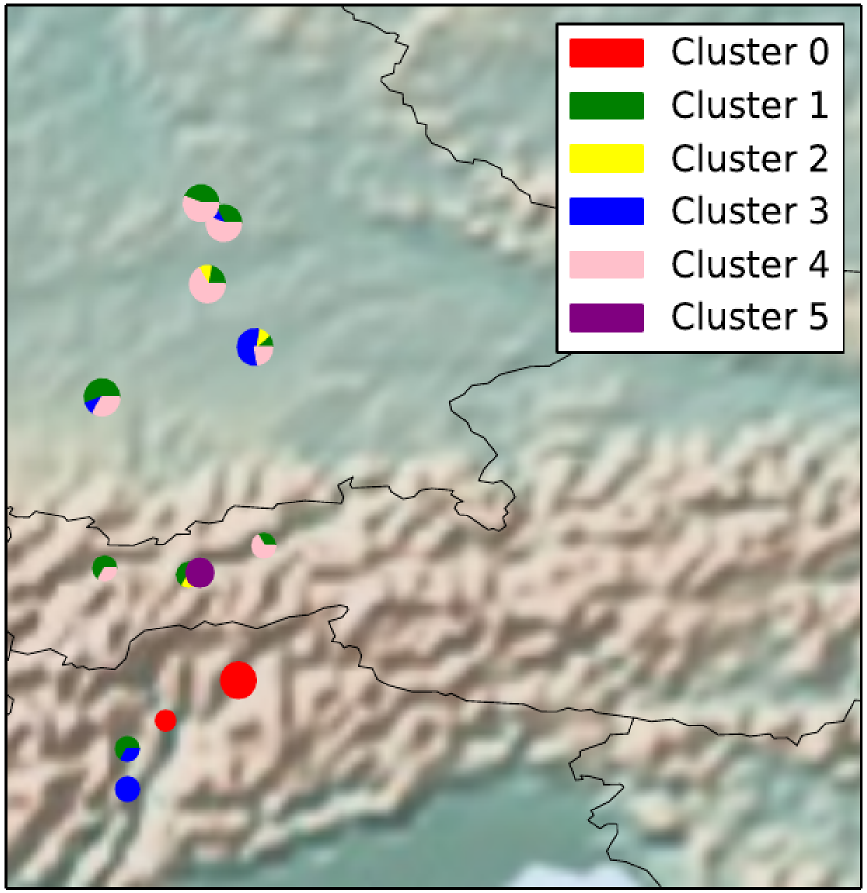
Unsupervised learning

How do the *clusters of isotopic-similar samples* correlate with the actual locations of the samples?

Settings

- All 7 isotope features used for clustering
- Assumption that data are generated by Gaussian mixture models
- EM algorithm to estimate the model parameters
- Cluster population

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



Detected clusters versus locations of the samples

Supervised learning

Are *region-specific models* good predictors for the origin of new samples?

Settings

- The data were categorized into classes “north”, “middle”, “south” Alps based on sample coordinates.
- 10-fold cross validation (9 folds for training, 1 for testing)
- A kNN classifier is built upon the training set
- The model is evaluated upon the test set

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0.833	0.115	0.837	0.833	0.832	0.868

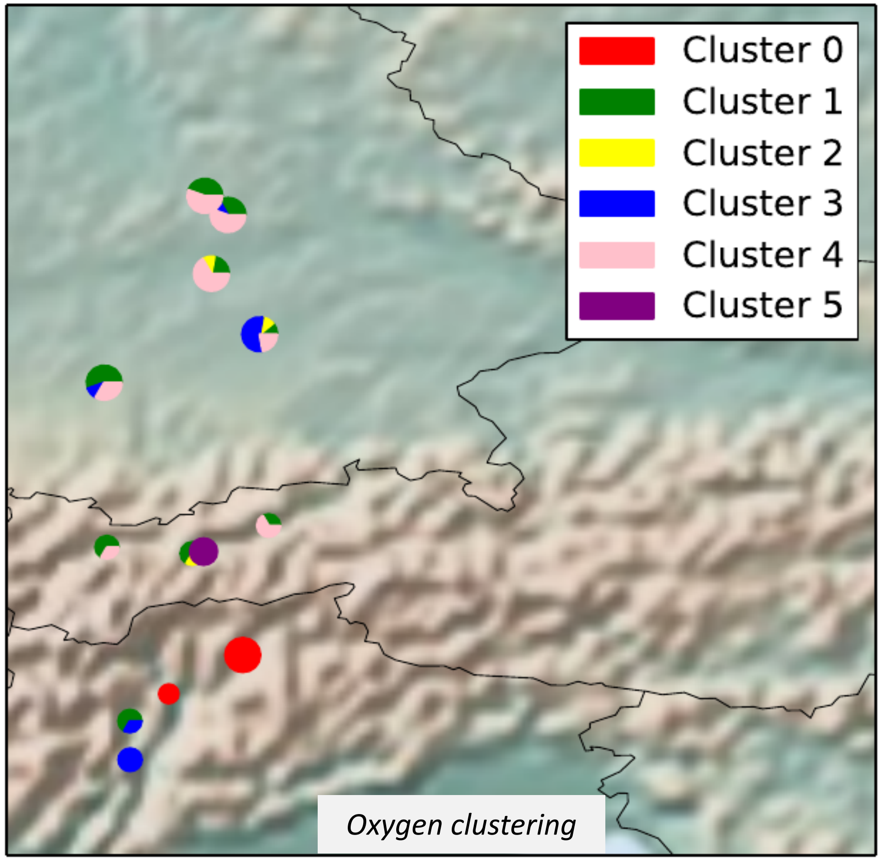
The effect of oxygen

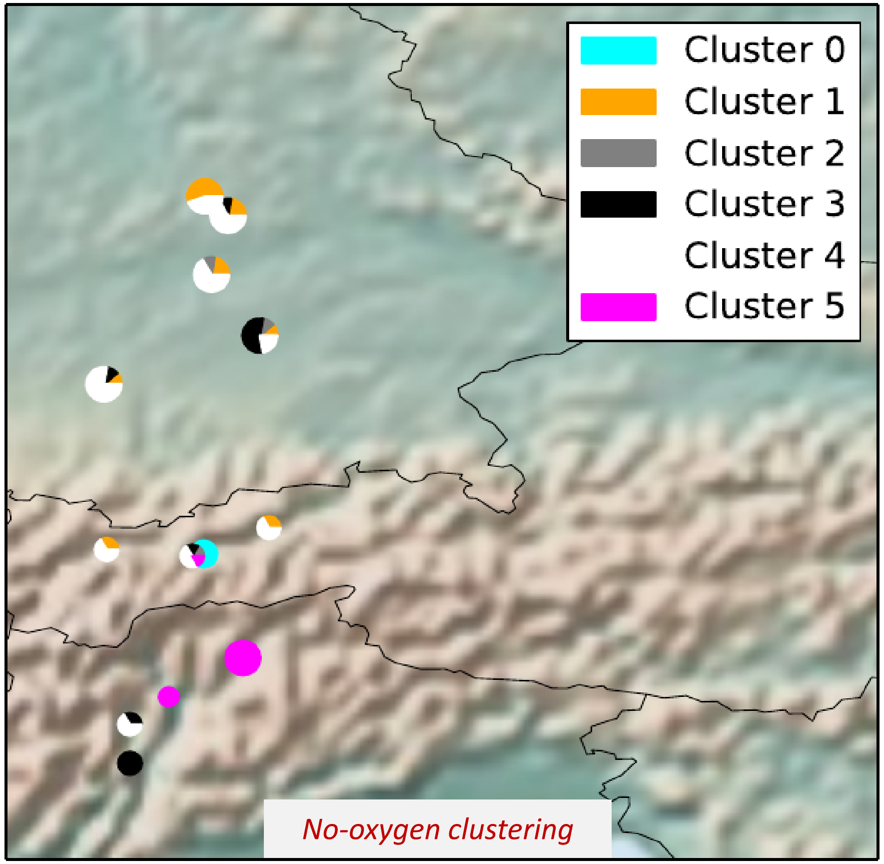
- The problem: Oxygen is sensitive to cremation, in contrast to strontium and lead

- Question: Is oxygen necessary for our analysis?
  - Quality might get lower of course but how worse?
- Why are interested in this?
  - A practical issue: we have a small uncremated sample set (~100 instances), it would be great if we can increase it by including uncremated samples.
  - A research question: how important is oxygen for fingerprinting?
  - A broader research question (for Data Mining): stability of data mining models under reduced feature spaces.

- Methodology:
  - Repeat the experiments by omitting oxygen
  - Find out how and where the with and without oxygen results "differ".
  - the ``differ" term depends on the Data Mining task per se

Unsupervised learning






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

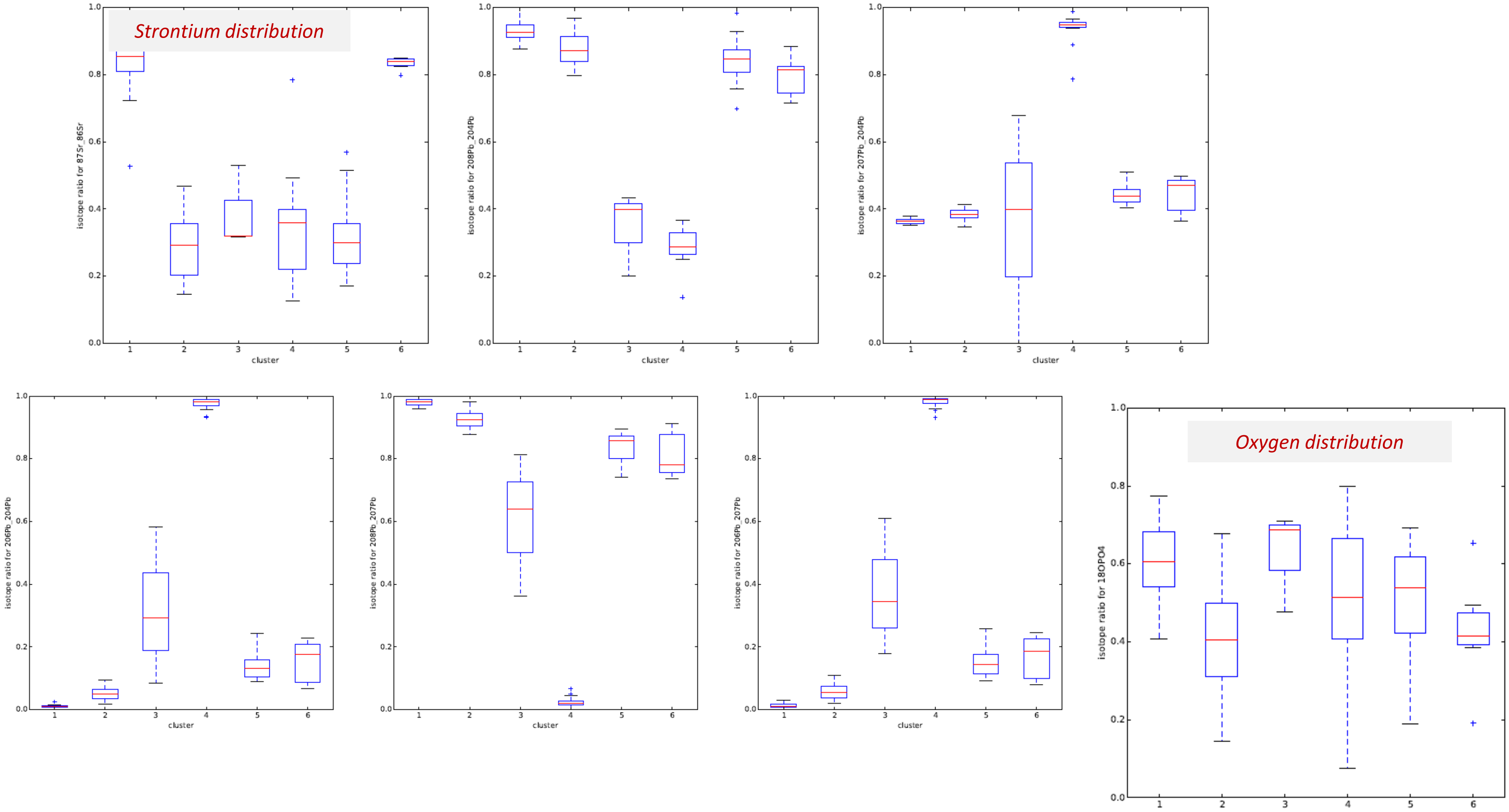
Detected clusters versus locations of the samples

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

Migration table

		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

Isotope distribution per cluster (Oxygen case)



Supervised learning

Are *region-specific models* good predictors for the origin of new samples?

Settings

- The data were categorized into classes “north”, “middle”, “south” Alps based on sample coordinates.
- 10-fold cross validation (9 folds for training, 1 for testing)
- A kNN classifier is built upon the training set
- The model is evaluated upon the test set

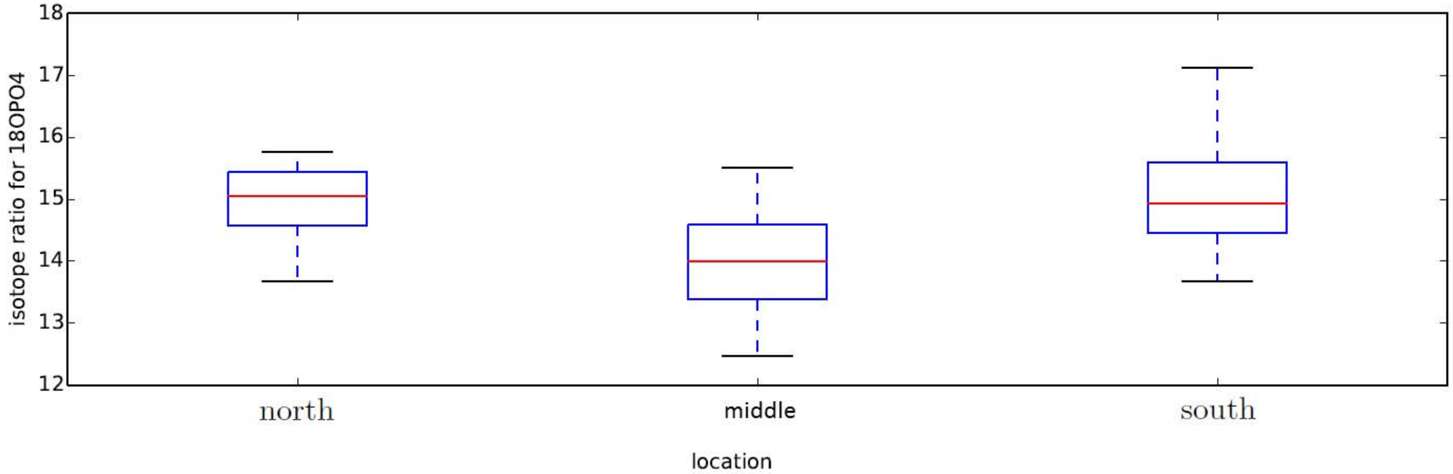
Evaluating the oxygen effect

- Experiments with and without oxygen

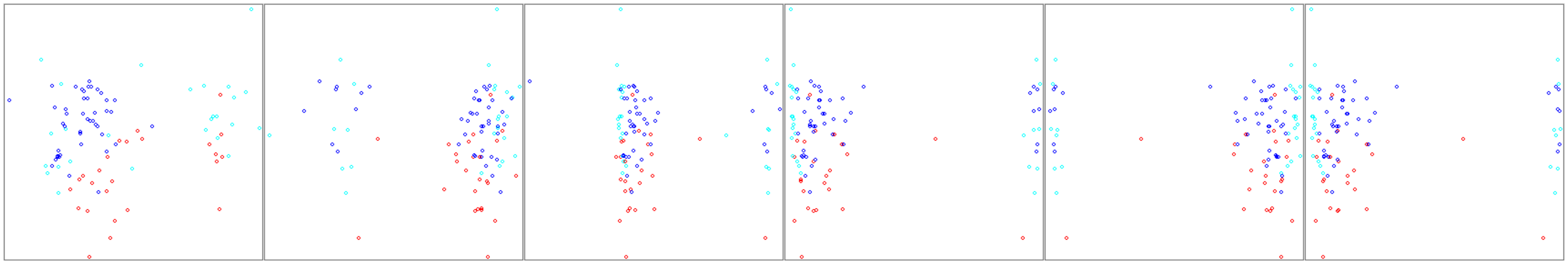
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Oxygen	0.833	0.115	0.837	0.833	0.832	0.868
No oxygen	0.76	1.168	0.768	0.76	0.759	0.785

How oxygen is correlated to other isotopes and location

- Oxygen isotope by location



- Oxygen correlation to other attributes



87Sr/86Sr | 208Pb/204Pb | 207Pb/204Pb | 206Pb/204Pb | 208Pb/207Pb | 206Pb 207Pb

Discussion on the findings and next steps

- Our sample is to small to make general statements
  - ~100 samples
  - Even less for the unsupervised case, since 10% is kept out for model testing
- Our initial analysis seems promising though
- Both supervised and unsupervised learning show that the omission of oxygen does not completely destroy the mining models, models are stable to a certain extend.
  - In the unsupervised case, most of the clusters of the oxygen case “survive” to the non-oxygen case.
  - In the supervised case, still acceptable performance scores
  - Lower scores are to be expected due to information loss incurred by oxygen omission

- A real crash test though would be the evaluation of models performance when the cremated samples are available.
- Combination of uncremated and cremated samples for model improvement.

Thank you for your attention

Questions?

More information on the technical report at the project's website:
htp:/www.for1670-transalpine.uni-muenchen.de