Application of Machine Learning Methods to Palaeoecological Data

[1]Marjeta Jeraj, [2]Sašo Džeroski, [2]Ljupčo Todorovski, [2]Marko Debeljak

[1] Department of Botany, University of Wisconsin-Madison, USA

[2] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

PALAEOECOLOGY

study of past environment

PALAEOENVIRONMENT

past climate

palaeofauna

palaeovegetation

humans

soils

lakes, rivers, oceans

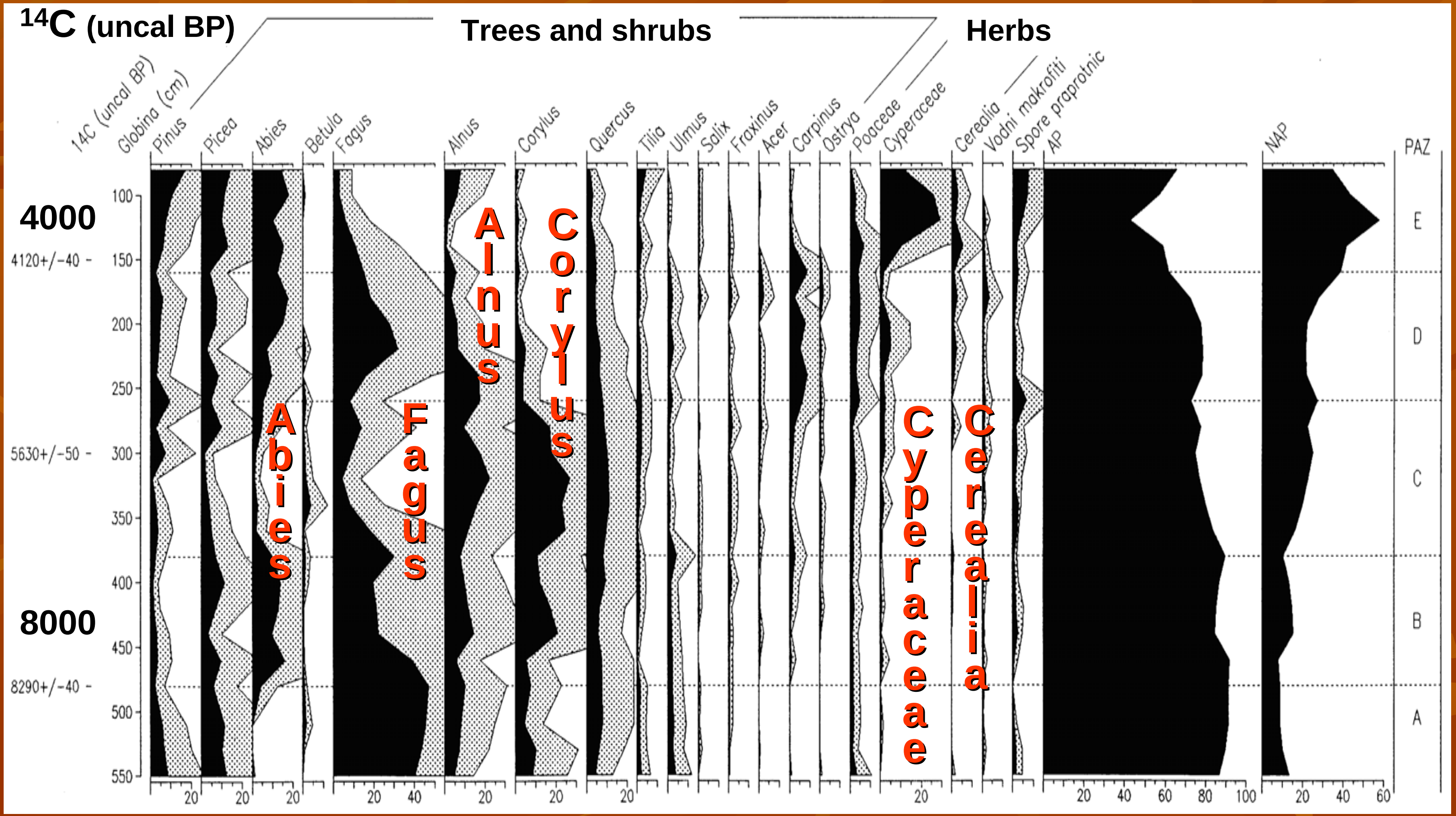
bedrock

Our study case:

Pollen data from SW Ljubljana Moor

Hočevarica and Bistra

Pollen diagram from Bistra core -
reconstruction of past vegetation dynamics



Relative pollen frequencies

Objectives

- find spatial and temporal correlations among different trees, shrubs and herbs,
  growing around SW Ljubljana Moor during the Early, Mid and Late Holocene periods

- find regularities and/or dependencies among co-existent plant species through time,
  with a focus on pile dwelling settlement period (app. 8400-3500 BP) to detect changes caused by humans

↓

application of machine learning techniques

to pollen data

Machine learning methods

1. Classification

2. Equation discovery

3. Hierarchical clustering

Classification

- predict a value of dependent variable from values of independent variables

- performed on pollen data from Bistra with WEKA j48 subpackage of classifiers, using J4.8 algorithm

- input: independent variables:

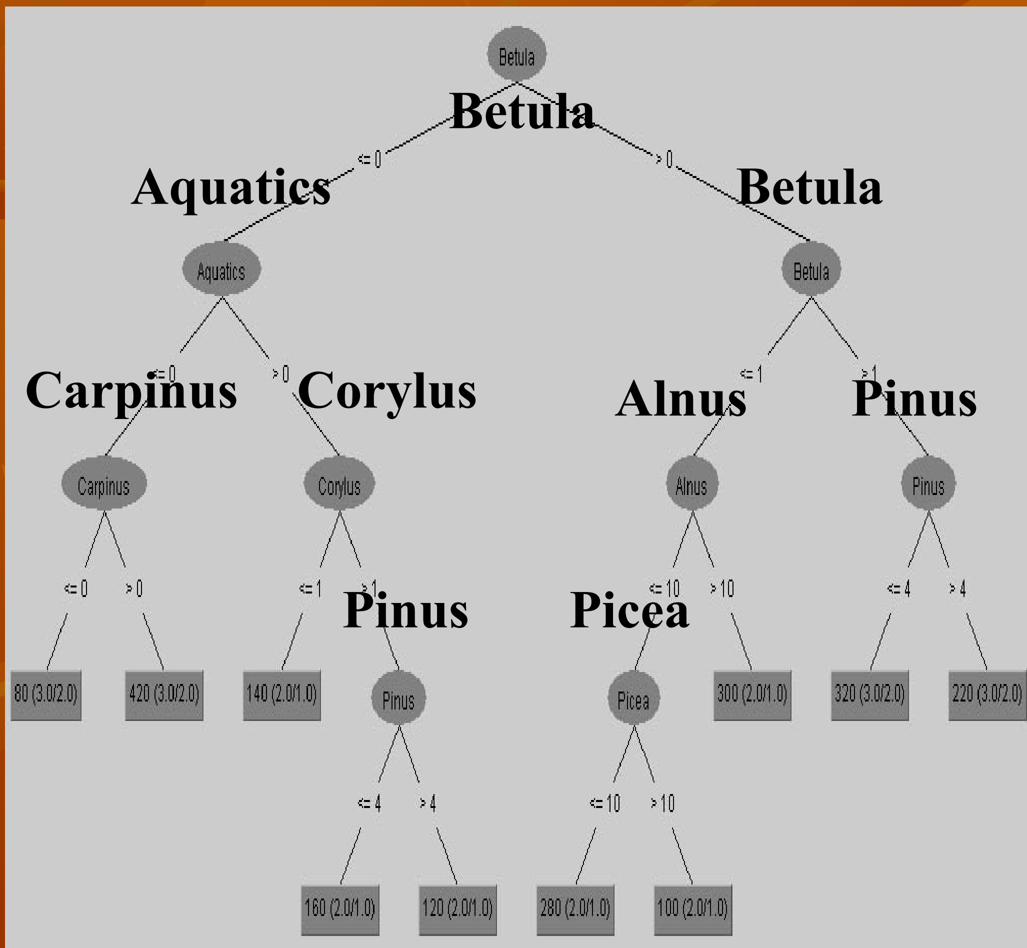
  - various depths between 80 and 550 cm, from where pollen samples were taken

  - relative pollen frequencies of Pinus, Picea, Abies, Betula, Fagus, Alnus, Corylus, Quercus, Tilia, Carpinus, other AP (arboreal pollen) and other NAP (non-arboreal pollen) at specific depths

- output: decision trees

Classification decision tree

from complete pollen dataset from Bistra



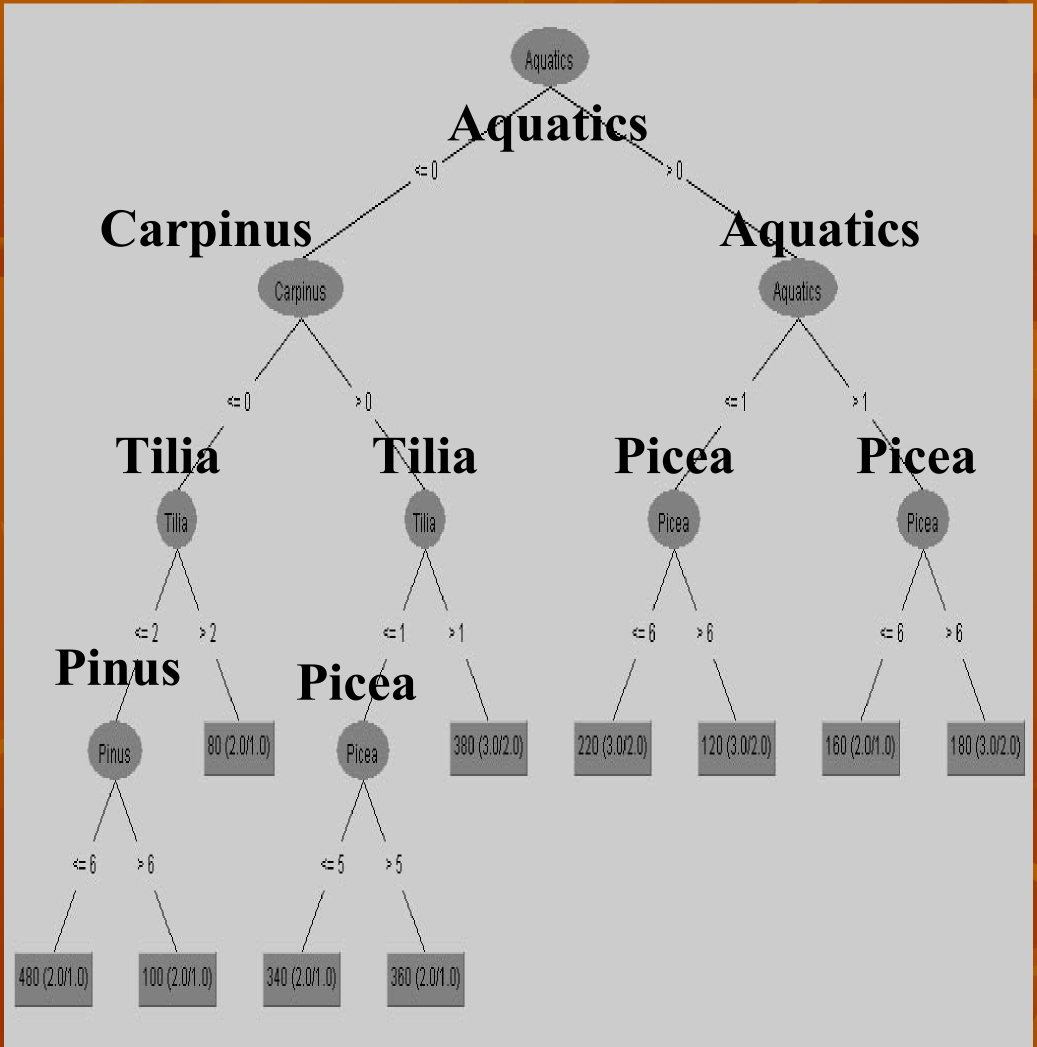
- Betula at the root → unexpected because of rare appearance in the pollen diagram;

  relevant because its presence appears binary

- discovered correlations differ from the ones found with pollen with the exception of Aquatics-Carpinus

Classification decision tree

from pollen dataset without Betula



- Aquatics at the root → also rare appearance and binary presence in the pollen diagram
- additional Tilia leaf appears, Corlyus and Alnus left out

Conclusions:

Classification predictive models **do not work well for pollen data** since they emphasize binary information and lower pollen values

Machine learning methods

1. Classification
2. Equation discovery
3. Hierarchical clustering

Equation discovery: LAGRANGE method

- discover differential equations,
  to determine the dynamics of a system from a time series of measured data
- input: a table of measured values
  our case: relative pollen frequencies from Bistra
- output: set of algebraic and ordinary differential equations, showing correlations between different plant genera/families and depth (age) over time

Equation discovery: LAGRANGE method

Correlations between plant types and depth from Bistra over time:

strongest correlations	**ecological applications for Lj. Moor**
Quercus = + 9.16 - 0.50 ***Pinus** (R = 0.76, S = 0.26)	Quercus: mesophilic, Holocene Pinus: cold-tolerant, Late Glacial
Pteridophytes = + 0.45 ***Pinus** (R = 0.82, S = 0.49)	preference to similar ie. acid soil type
Depth = +149.2 + 7.6 ***Fagus** (R = 0.72, S = 0.32)	pioneer Early Holocene species, afterwards competition, human impact

Equation discovery: LAGRANGE method

strongest correlations	Ecological applications for Lj. Moor
Quercus = + 4.1 + 0.24 ***Corylus** (R = 0.80, S = 0.24)	tolerant to similar conditions, wood used by pile dwellers
Aquatics = + 0.28 + 0.27 ***Carpinus** (R = 0.76, S = 0.66)	both prefer bright and open areas, e.g. after initial forest clearance
Cerealia = + 0.30 + 0.12 ***Cyperaceae** (R = 0.85, S = 0.48)	indicators for human disturbance and changes in water level
Depth = +394.3 - 113.1***Cerealia** (R = 0.81, S = 0.27)	traces of early cultivation in upper layers

Machine learning methods

1. Classification

2. Equation discovery

3. Hierarchical clustering

Hierarchical clustering

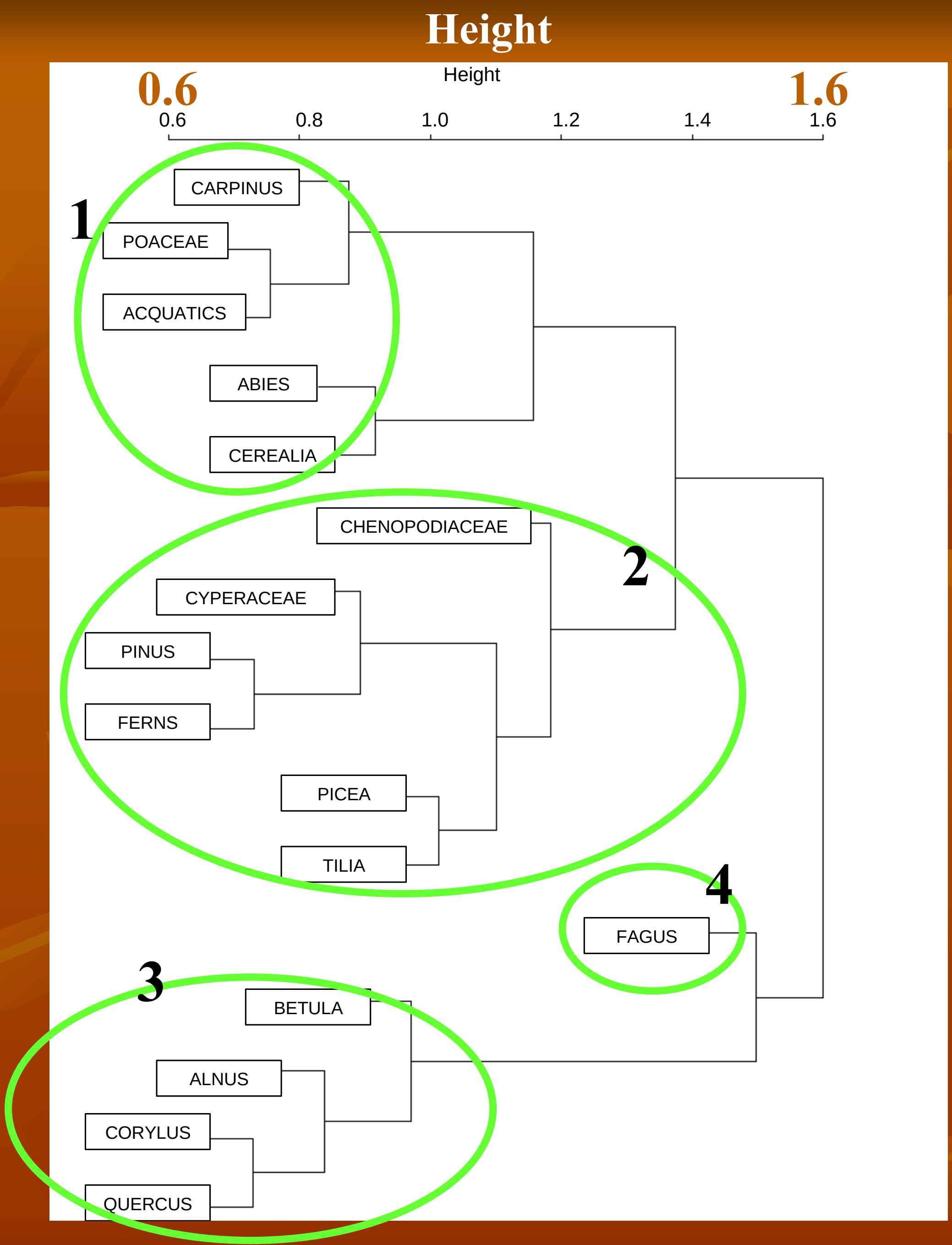
- similarity-based method, which calculate a suitable distance between time series (x,y) using correlation coefficient (R)

$$d(x, y) = \sqrt{2(1 - R_{x,y})}$$

- input: distances (d) between time series
  our case: between plant genera/species from Bistra and Hočevarica

- output: dendrogram

- appropriate for short time series



Cluster dendrogram for pollen data from Bistra

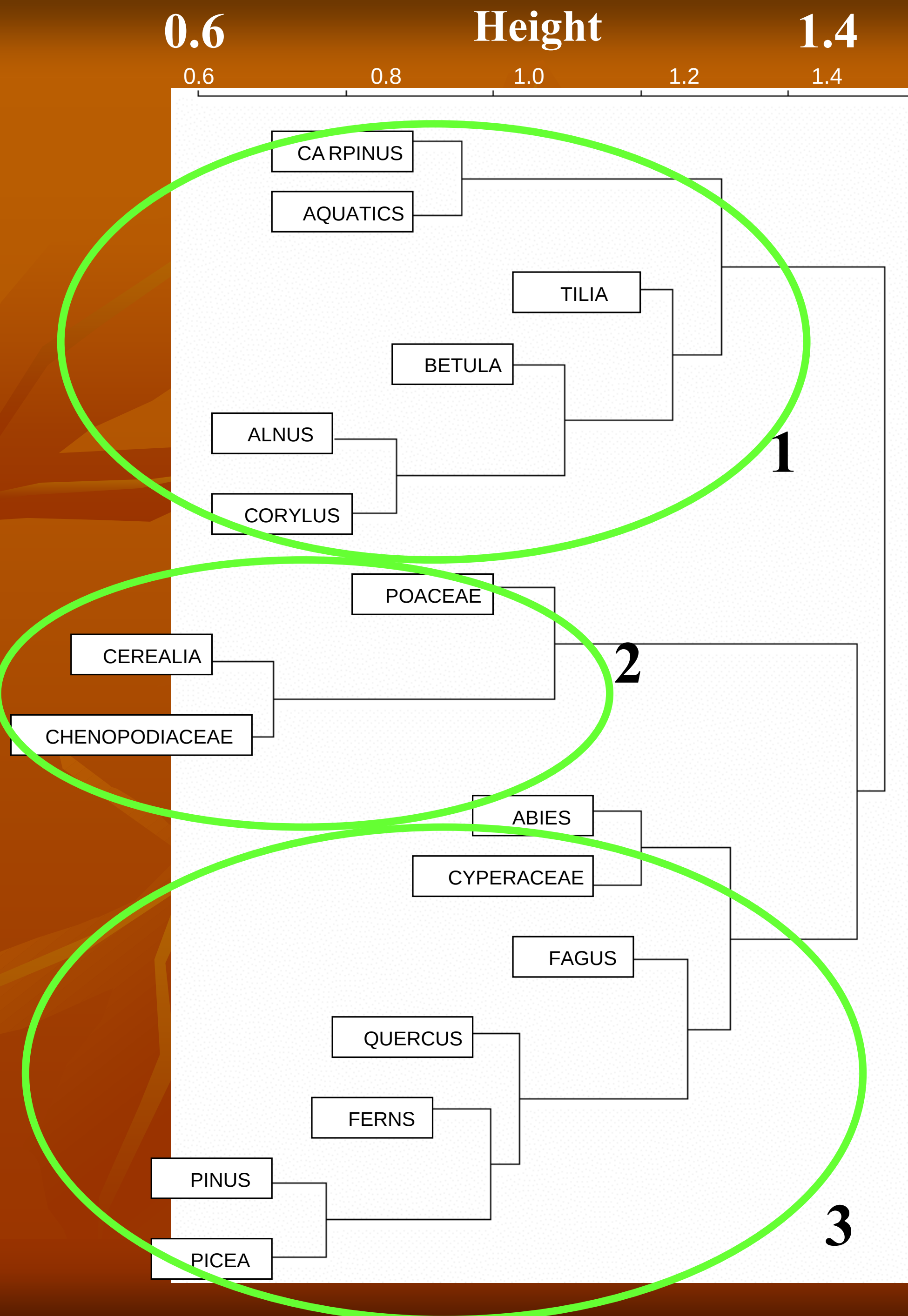
- four groups of **clusters**:

  max. appearance

  1 Late Holocene
  2 heterogeneous
  3 Mid Holocene
  4 Early Holocene

- **strongest correlations** (0.6-0.8):

  in 1: aquatics-Poaceae-Carpinus
  in 2: Pinus - ferns
  in 3: Corylus - Quercus



Cluster dendrogram for pollen data from Hočevarica

- three groups of **clusters**:

  max. appearance

  1 Late Holocene
  2 heterogenous
  3 some: Early Holocene

- **strongest correlations** (0.6-0.8):

  in 1: Alnus - Corylus
  in 2: Cerealia - Chenopodiaceae
  in 3: Pinus - Picea

Synthesis of results

Equation discovery and hierarchical clustering: the most significant results

strongest correlations	LAGRANGE (R)	Clusters (d): Bistra	HO
Cerealia-Cyperaceae	+ 85		
Cerealia-Chenopodiaceae			0.7
ferns-Pinus	+ 82	0.7	
(ferns-Pinus)-Cyperaceae		0.9	
Pinus-Picea			0.8
(Pinus-Picea)-ferns			1.0
depth-Cerealia	- 81		
Quercus-Corylus	+ 80	0.7	
(Quercus-Corylus)-Alnus		0.9	
Alnus-Corylus			0.9
Aquatics-Poaceae		0.7	
Aquatics(-Poaceae)-Carpinus	+ 76	0.9	

Conclusions

- machine learning methods applied to pollen data from from Bistra and Hočevarica, especially **equation discovery** and **hierarchical clustering**, successfully detected different patterns in historical compositions of plant communities

↓

they proved to be a **good predictor of correlations among different plant species/genera** growing around SW Lj. Moor in the past

- some of the discovered correlations can be well **interpreted with** the existing knowledge about **ecological** relations among plant species, as well as with available **archaeological and archaeobotanical information**

Conclusions

- **some** of the existing **relationships** among plants are **not detected** with any of the applied machine learning approaches;

  on the other hand, they **detected** some correlations that are **not obvious** from pollen diagrams

- in any case, interpretations of relationships in composition and dynamics of historical plant communities, discovered by machine learning techniques, require sufficient **ecological and interdisciplinary background**, and collaboration between palaeoecologists and statisticians

Thank you for your attention