Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecological Modelling 191 (2006) 159–169
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Application of machine learning methods to palaeoecological data

Marjeta Jeraj [a,*], Sašo Džeroski [b], Ljupčo Todorovski [b], Marko Debeljak [b]

[a] *Department of Botany, University of Wisconsin, Madison, WI, USA*
[b] *Department of Knowledge Technologies, "Jožef Stefan" Institute, Ljubljana, Slovenia*

Available online 19 September 2005

Abstract

A palaeoecological study was conducted to investigate past environmental conditions and vegetation dynamics around the southwestern Ljubljana Moor. In order to find potential regularities and/or dependencies among co-existent plant species through time, different machine learning methods were applied to pollen records from the cores taken at Bistra and Hočevarica. The data comprised relative pollen frequencies of the most common plant genera/families at particular core depths that correspond to particular ages in the Early and Mid Holocene periods. The applied methods include equation discovery and hierarchical clustering. Both methods have found plausible and explainable relationships among identified plant genera/families.

Keywords: Palaeoecology; Vegetation dynamics; Machine learning; Equation discovery; Hierarchical clustering

1. Introduction

Palaeoecology is the study of past environment, including palaeoclimate, geomorphology, past hydrology, soil development, palaeovegetation, palaeofauna, and human settlement history (Bell and Walker, 1992). It provides a reconstruction of past environmental conditions and dynamics, and allows insight into their causes and relationships. Palaeoecological material belongs to different geological periods, ranging from before the Cambrian to recent centuries. From a human history point of view, the Quaternary period and especially the Holocene are the most interesting, since they reflect long-term interactions between people and their environments (Birks and Birks, 1980; Lowe and Walker, 1997).

Several machine learning approaches, especially predictive modeling (i.e., classification and regression) and clustering have been successfully applied to various ecological studies (Džeroski et al., 1999; Bratko et al., 2003). Moreover, methods such as regression, nearest neighbor, and hierarchical Bayesian modeling have been proven to provide good estimates for palaeoecological reconstructions (Manilla et al., 1998). In our analyses, palynological data, which enable us to reconstruct past vegetation on the basis of pollen, were used and analyzed with machine learning methods.

Our goal in this paper is to find dependencies among co-existent plant species, using machine learning tools. More specifically, we are looking for temporal correlations among different trees, shrubs,

* Corresponding author.
E-mail address: mjeraj@wisc.edu (M. Jeraj).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.08.018

and herbs, which were growing around the southwestern Ljubljana Moor during the last 9000 years. We investigate correlations between a particular plant taxon and a depth of the core from where its pollen was found. We are especially interested in the Early and Mid Holocene periods, from about 8400 until 4000 BP, to compare the environmental settings before, during, and after the initial occupation. In addition, we attempt to detect changes in vegetation caused by the appearance of humans. We investigate relationships among plant genera/families significant for particular plant communities. We also wonder whether the presence of any particular plant type was correlated with the absence of another, and whether there were any correlations with anthropogenic indicators.

Table 1
The relationship between depth and chronology for the dataset from Bistra

Depth (cm)	^{14}C age		Period
	Uncal BP	Cal BP	
100	3800	4100	Late Holocene
150	4100	4650	
200	4600	5150	Mid Holocene
250	5150	5800	
300	5600	6400	
350	6400	7200	
400	7100	8000	Early Holocene
450	7800	8800	
500	8600	9600	

2. Data and methods

We apply two machine learning methods, equation discovery and hierarchical clustering to the pollen data from the cores taken on the southwestern Ljubljana Moor in Slovenia (Jeraj, 2004). The pollen datasets derive from chemically processed sediment samples that were obtained from a 550 cm deep core at Bistra and from a 420 cm deep core at Hočevarica. The sampling frequency in the cores was every 10 cm. The upper layers below the surface were not suitable for pollen analyses, because of pedogenesis and contamination. The analyzed data, which are shown in the pollen diagrams (Fig. 1), consist of relative pollen frequencies of the most common plant genera/families at particular depths that correspond to particular chronological sequences. On the basis of three radiocarbon dates, available for the dataset from Bistra, we were able to establish the relationship between depth and chronology, shown in Table 1. The main events in vegetation history match in the pollen diagrams from both sites, which are located in the close vicinity and thus capture the pollen from the same region.

Relative pollen frequencies in the diagrams refer to percentages of the pollen of individual taxon in relation to the total pollen sum in the sample. In each diagram different plant taxa and types are grouped into several categories, including trees and shrubs, herbs, aquatics, and spores (Pteridophytes). The rightmost vertical graphs show the ratio between arboreal (AP) and non-arboreal (NAP) pollen. The pollen diagrams are divided into several pollen assemblage zones (PAZ) according to major changes in pollen percentages through time. The radiocarbon dates on the left of the diagrams (^{14}C) provide chronological estimates for inferred vegetation and environmental events.

The machine learning methods that we used for paleodata from southwestern Ljubljana Moor are further described in more details, and their results are compared and discussed.

In the first series of experiments, we use equation discovery method LAGRANGE (Džeroski and Todorovski, 1995) to identify positive and negative correlations between pairs of genera/families concentrations. Equation discovery methods deal with the task of automated discovery of quantitative laws and models, expressed in the form of equations, in collections of measured data. Given a table of measured values of the observed variables, equation discovery methods try to find one or more equations that minimize the discrepancy between the measured values of the system variables and values predicted using the equations. LAGRANGE is capable of discovering polynomial equations with limited degree and length and uses *multiple correlation coefficient* (also referred to as *coefficient of determination*) as a measure of agreement between the measured and predicted values of the system variables. Maximal degree and length of the polynomial equations as well as the minimal correlation coefficient thresholds are user-defined parameters. In the experiments presented here, we use LAGRANGE to discover linear (i.e., first degree) equations that relate pairs of genera/families concentrations (i.e., minimal

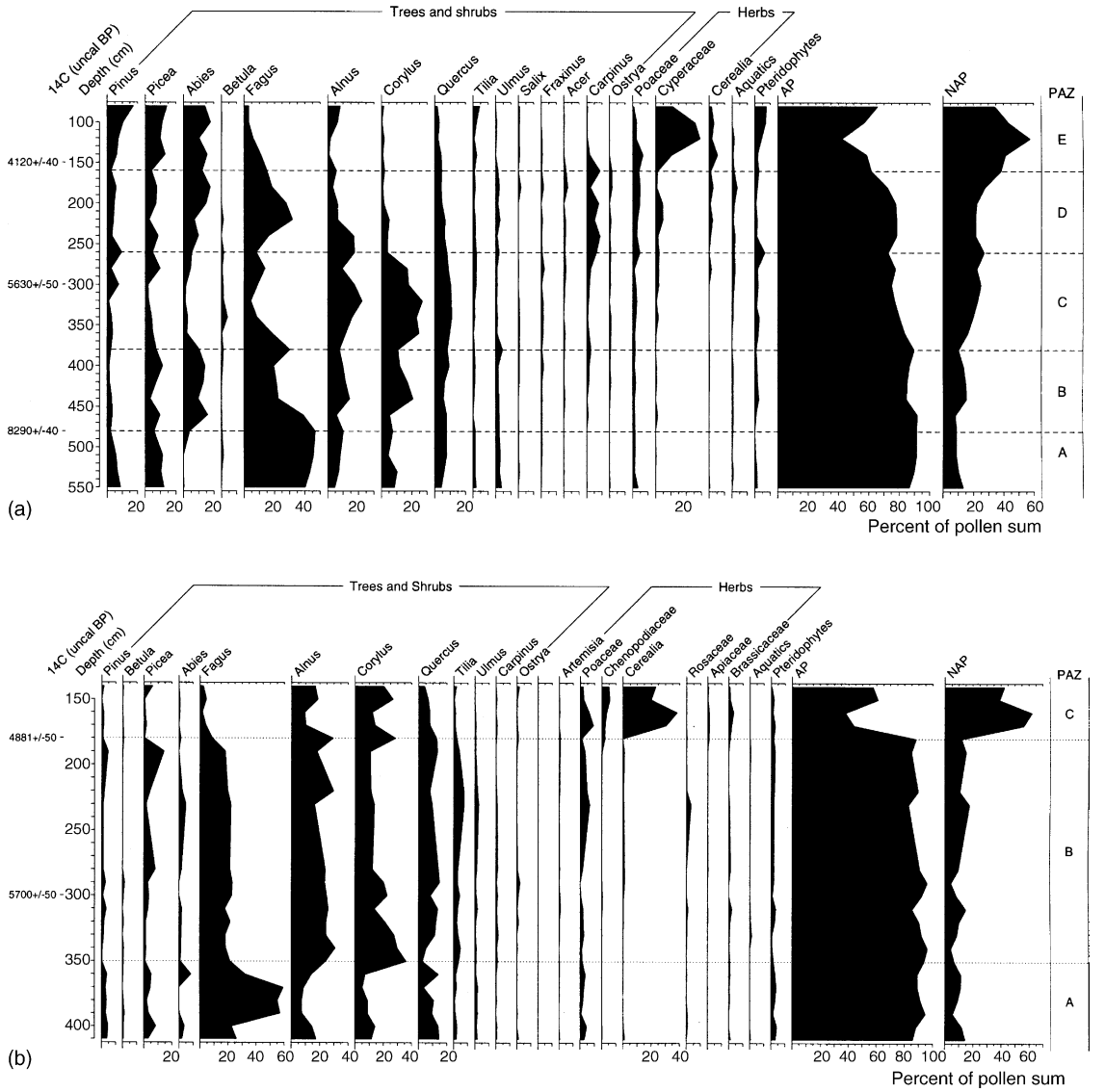

Fig. 1. Pollen diagrams from Bistra (a) and Hočevarica (b) cores. The axis x presents relative pollen frequencies (%) for particular genus/family and the axis y refers to different depths which correspond to different ages.

equation length) and have multiple correlation coefficient of at least 0.7.

In the second series of experiments, we used a method for hierarchical clustering of time series data (Todorovski et al., 2002) to identify groups of positively correlated concentrations of genera/families. In general, clustering is an unsupervised learning method that from data about a number of objects (in our case, genera/families) creates groups of objects following two criteria. First, objects are close (or similar) to the other objects from the same group (internal cohesion) and distant (or dissimilar) from objects in the other groups (external isolation). Since our objects are represented using time series of concentrations, we used correlation coefficient to measure the similarity. In particular the distance between two genera/families g_1 and g_2 is being measured as $d(g_1, g_2) = \sqrt{1 - r(g_1, g_2)}$, where $r(g_1, g_2)$ is the correlation coefficient between

time series for g_1 and g_2. A particular class of clustering methods, widely used in statistical data analysis, are hierarchical clustering methods. The hierarchical clustering algorithm starts with assigning each object to its own cluster, and iteratively joins together the two closest (most similar) clusters together. The distances between objects are provided as input to the clustering algorithm. The iteration continues until all objects are clustered into a single cluster. We applied the complete linkage method to calculate distances between joined clusters. The output of a hierarchical clustering algorithm is a hierarchical tree or dendrogram, where the height of each node is proportional to the distance between the joined clusters.

3. Results and interpretation

3.1. Equation discovery with LAGRANGE

Table 2 presents the results of applying LAGRANGE to the Bistra dataset. We used LAGRANGE to discover linear equations with multiple correlation coefficient above or equal to 0.70. Each of them shows the correlation between a pair of genera/families. The equations marked in bold can be adequately interpreted from ecological point of view and possible explanations are listed below. In general, these are equations with higher multiple correlation coefficients (typically above 0.75), although there are few exceptions, which include non-frequent species, e.g. birch (Betula).

Table 2
The results of the LAGRANGE equation discovery analysis, applied to the pollen dataset from Bistra

Discovered equation	Multiple correlation coefficient (R)
Quercus = +9.16 − 0.50 × Pinus	**0.76**
Tilia = +0.89 + 0.19 × Pinus	0.70
Pteridophytes = +0.45 × Pinus	**0.82**
Tilia = +0.23 × Picea	0.74
Corylus = +6.3 + 9.8 × Betula	0.70
Quercus = +5 + 3.5 × Betula	0.83
Other AP = +0.43 + 1.3 × Betula	0.71
Depth = +149.2 + 7.6 × Fagus	**0.72**
Quercus = +4.1 + 0.24 × Corylus	**0.80**
Other AP = −1.15 + 0.33 × Quercus	0.75
Poaceae = +1.79 + 0.24 × Carpinus	0.72
Aquatics = +0.28 + 0.27 × Carpinus	**0.76**
Cerealia = +0.30 + 0.12 × Cyperaceae	**0.85**
Depth = +394.3 − 113.1 × Cerealia	**0.81**

Palaeoecological explanations for the correlations in Table 2 (marked in bold) are as follows:

- Oak (Quercus) and pine (Pinus) show a negative correlation ($R = 0.76$), which is expected because pines prefer cold conditions, whereas the oaks growing around Ljubljana Moor (*Q. robur*, *Q. petraea*) are typical mesophilic trees. This is also supported by the fact that in the regional forest succession pine used to be one of the dominant tree species during the Late Glacial, while oak appeared among the most common species as the climate began to warm in the Early Holocene.
- Ferns (Pteridophytes) and pine (Pinus) have a strong positive correlation ($R = 0.82$). This can be explained by their preference for similar soil types; ferns usually grow in understory on acidic soils, which are favorable for pines as well.
- Even stronger positive correlation of beech (Fagus) with depth ($R = 0.72$) would be expected since beech appeared to be one of the pioneer species entering Holocene forests after the glaciation. Later, it was exposed to competition with other species and cut for logging.
- Oak (Quercus) and hazel (Corylus) express a positive correlation ($R = 0.80$), which may be ascribed to their similar tolerance to climatic conditions. During the pile dwelling period in Mid Holocene they both became less frequent, although we are not able to assign this directly to influences by pile dwellers.
- Aquatic plants (Aquatics) and hornbeam (Carpinus) are also positively correlated ($R = 0.76$), possibly because they both prefer more open and bright areas, which emerged around Ljubljana Moor after initial forest clearance.
- Cereals (Cerealia) and sedges (Cyperaceae) show the strongest correlation ($R = 0.85$). They both indicate human appearance and activities in the area, largely associated with agriculture and forest clearing. In addition, they also imply considerable changes in water level in the area.
- Strong negative correlation of cereals (Cerealia) with depth ($R = 0.81$) is very reasonable and expected. Cereal pollen was only found in shallower sediments, dated about 5600 years BP, when first set-

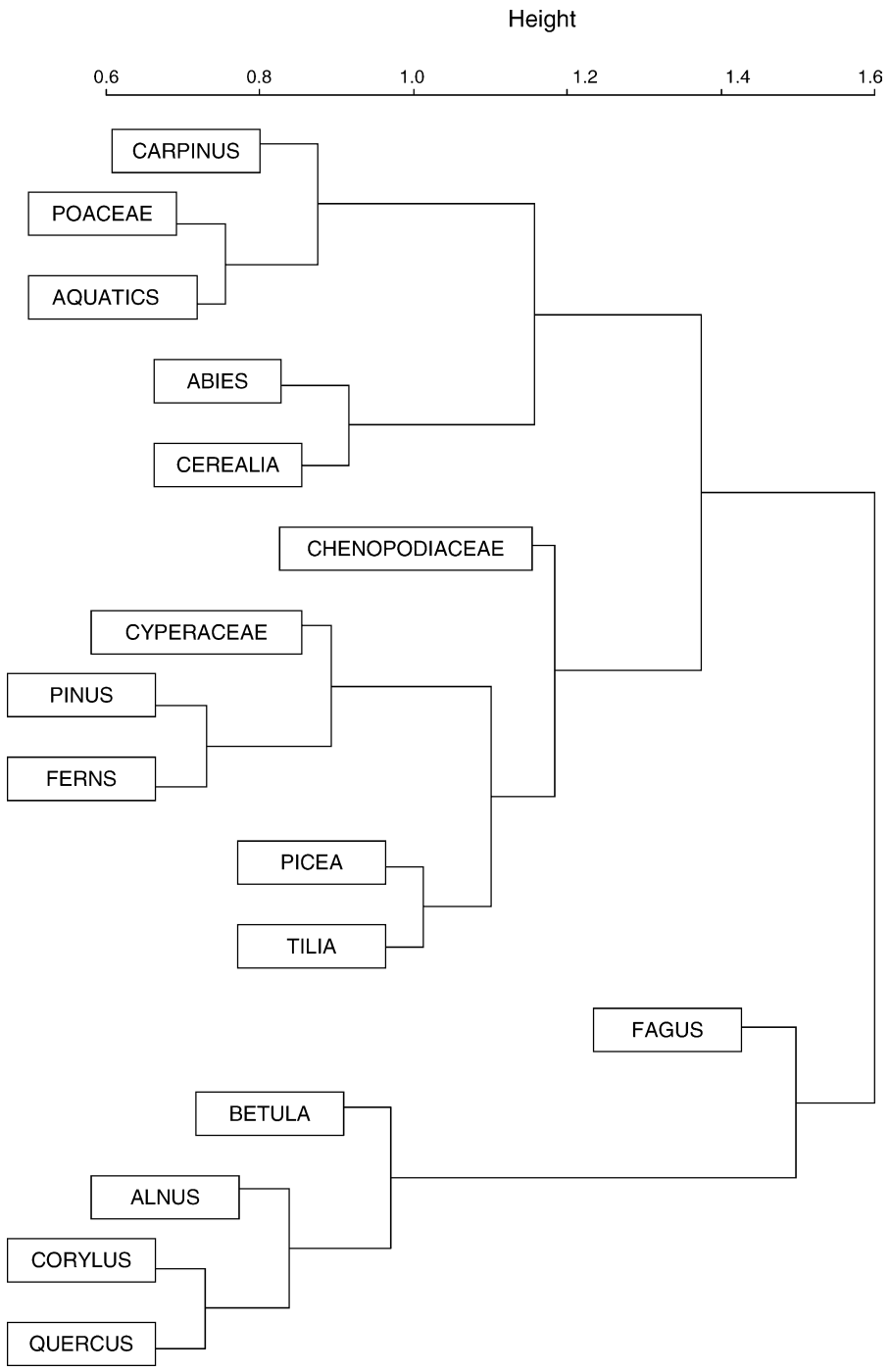

Fig. 2. Cluster dendrogram for the pollen dataset from Bistra. Heights refer to the distance between various plant genera/families or groups.

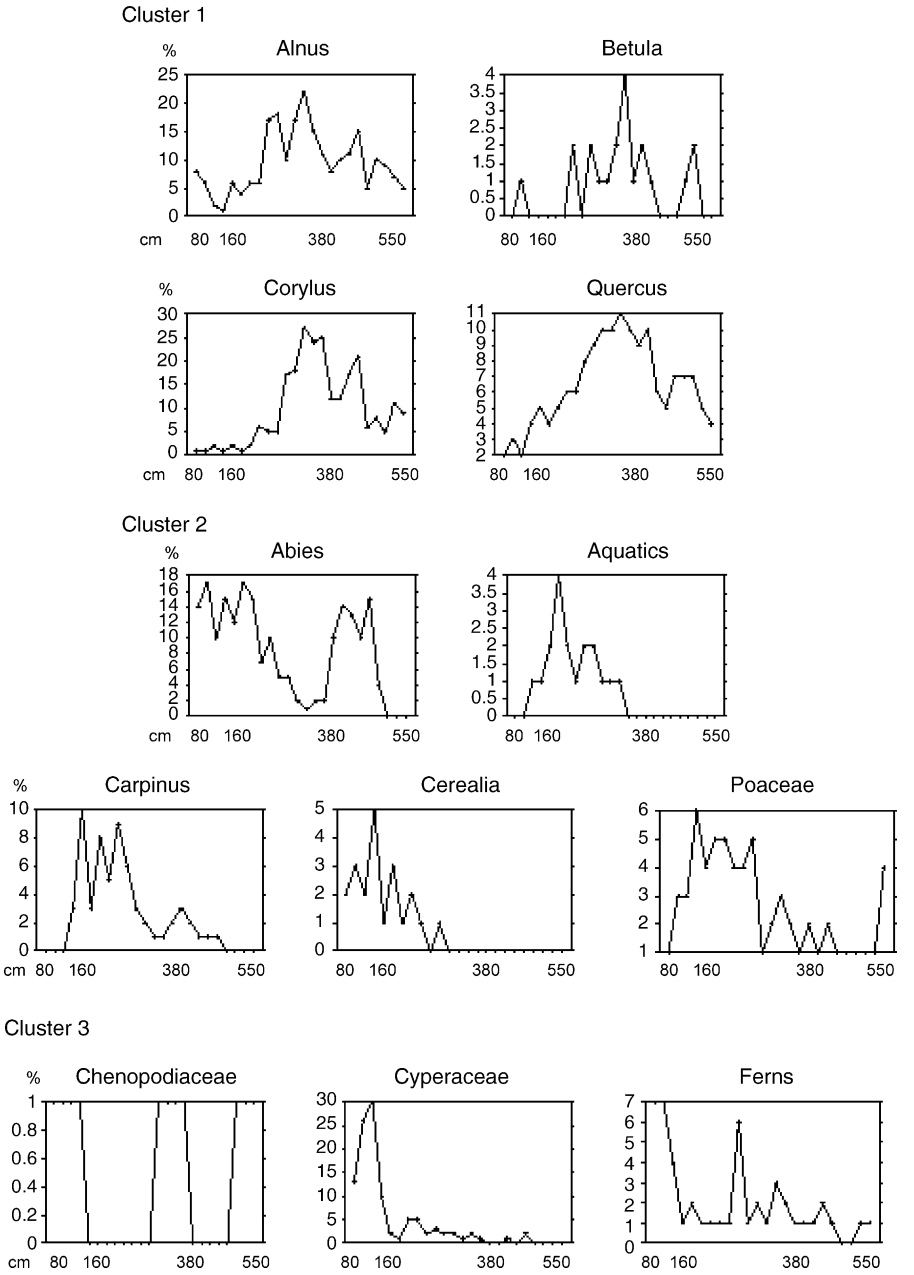

Fig. 3. The four clusters of genera/families identified for the Bistra dataset. On the axis *y* relative pollen frequencies (%) for particular genus/family are presented and the axis *x* refers to different depths which correspond to different times. Depths between 80 and 160 cm represent a part of the Late Holocene and the beginning of the Mid Holocene (including the settlement period), depths between 160 and 380 cm refer to the rest of the Mid Holocene, and depths between 380 and 550 cm to the Early Holocene period.

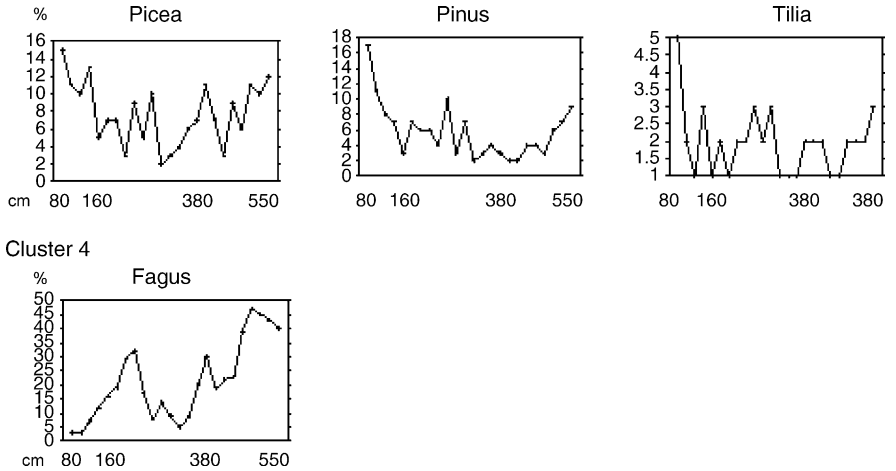

Fig. 3. (*Continued*).

tlers occupied the area and started cultivation (Jeraj, 2002).

3.2. Hierarchical clustering of time series data

Hierarchical clustering was performed on both Bistra and Hočevarica datasets. In the case of Bistra, the cluster dendrogram, presented in Fig. 2, comprises four distinct clusters: cluster 1 with oak, hazel, alder, and birch (Quercus, Corylus, Alnus, and Betula), cluster 2 with grasses, aquatic plants, hornbeam, fir, and cereals (Poaceae, Aquatics, Carpinus, Abies, and Cerealia), cluster 3 with pine, ferns, sedges, spruce, linden, and orachs (Pinus, Pteridophytes, Cyperaceae, Picea, Tilia, and Chenopodiaceae), and cluster 4 with beech (Fagus). Fig. 3 presents the time series for each cluster. The species from cluster 1 reached their maximum appearance in the middle of the observed period (in the Mid Holocene), while the species from cluster 2 appeared with their highest abundance in the Late Holocene after human occupation. The appearance of species from cluster 3 is more heterogeneous, and the appearance of beech (Fagus) in cluster 4 generally increases with time. Note also that species from cluster 1 are the most similar and cluster 4 shows the lowest correlation with other pollen types.

In the cluster dendrogram from Bistra distances are calculated on the basis of correlations among different plant genera/families, and range between 0.6 and 1.6 (Fig. 2). From the ecological point of view, the correlations that correspond to small distance values are more plausible, either within clusters or among them.

A distance between 0.6 and 0.8 corresponds to plant genera/families or groups with the most similar patterns and changes over time. These are pines (Pinus) and ferns (Pteridophytes), grasses (Poaceae) and aquatic plants (Aquatics), and hazel (Corylus) and oak (Quercus). For two of the listed pairs, Pinus/ferns and Corylus/Quercus, a strong correlation was also found with the equation discovery analysis and it is already discussed in the previous subsection. In the cluster dendrogram, grasses (Poaceae) and aquatic plants (Aquatics) are further strongly correlated with hornbeam (Carpinus), which might be due to their similar preference to more open landscape with sufficient amount of light. A part of this correlation, for aquatic plants and Carpinus, was also demonstrated by the equation discovery method.

Distances between 0.8 and 1.0 can be still interpreted as strongly correlated clusters and the following groups can be identified: (a) Pinus/ferns and Cyperaceae, (b) Corylus/Quercus and Alnus, and (c) Abies and Cerealia (Fig. 2). Pine (Pinus), ferns and sedges (Cyperaceae) appeared to be among the dominant vegetation types after the initial occupation and may indicate changes in the surrounding environment such as the establishment of a marshy landscape with patches of degraded soil, most suitable for pioneer species. The pattern of co-appearance of oak (Quercus), hazel (Corylus), and alder (Alnus) seems to be rather com-

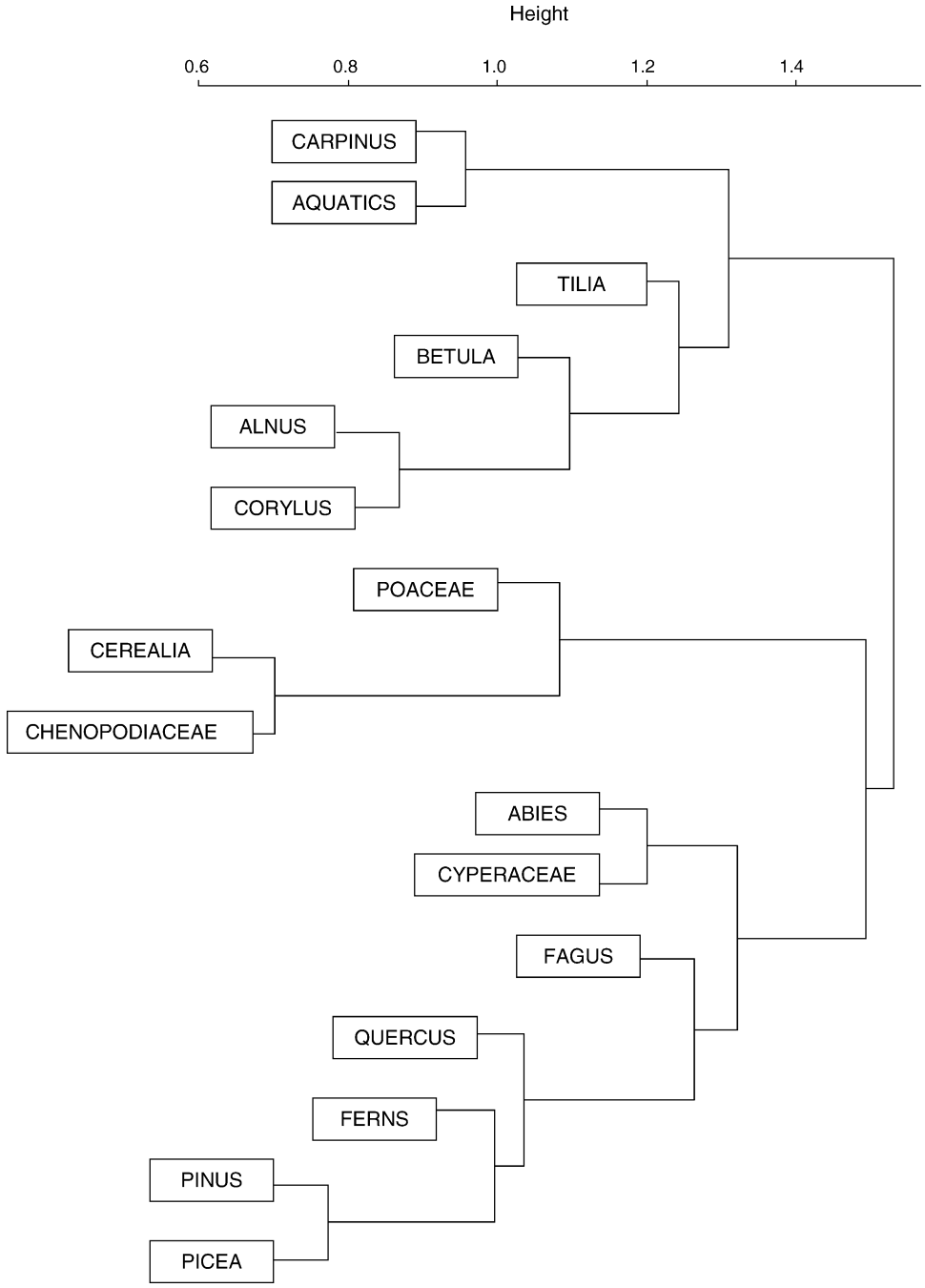

Fig. 4. Cluster dendrogram for the pollen dataset from Hočevarica. Heights refer to the distance between various plant genera/families or groups.

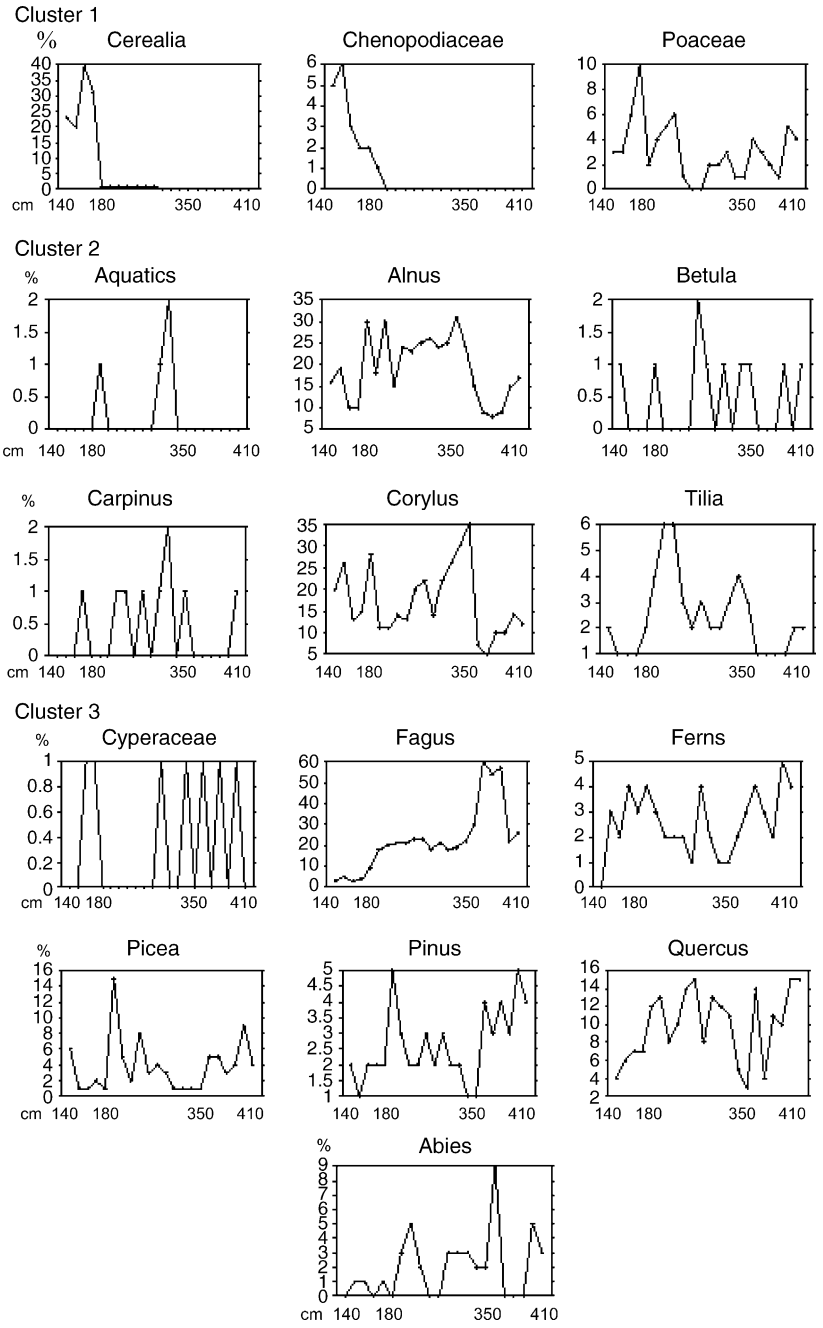

Fig. 5. The three clusters of genera/families identified for the Hočevarica dataset. On the axis *y* relative pollen frequencies (%) for particular genus/family are presented and the axis *x* refers to different depths which correspond to different times. Depths between 80 and 160 cm represent a part of the Late Holocene and the beginning of the Mid Holocene (including the settlement period), depths between 160 and 380 cm refer to the rest of the Mid Holocene, and depths between 380 and 550 cm to the Early Holocene period.

mon in the area, especially in the later Holocene periods when the succession of primary forest phases ended and mesophilic trees started to appear together as secondary phases. For the cluster with fir (Abies) and cereals (Cerelia), we were unable to give a reasonable ecological interpretation of their correlation.

In the cluster dendrogram for the Hočevarica dataset, presented in Fig. 4, three clusters can be distinguished. Cluster 1 is composed of cereals (Cerealia), orachs (Chenopodiaceae), and grasses (Poaceae), which have their maximum values in the Late Holocene during the settlement period (Fig. 5). Cluster 2 includes hazel (Corylus), alder (Alnus), aquatic plants (Aquatics), hornbeam (Carpinus), birch (Betula), and linden (Tilia), with rather heterogeneous temporal behavior. In cluster 3, pine (Pinus), spruce (Picea), ferns (Pteridophytes), oak (Quercus), fir (Abies), sedges (Cyperaceae), and beech (Fagus) are grouped, and some of them show a tendency of increasing frequency with time. Similarly, as for the Bistra dataset, distance measures between pollen types range between 0.6 and 1.6. The strongest correlation, i.e., the lowest distance is shown for Cerealia and Chenopodiaceae.

This analysis supports the evidence for strong co-appearance of cereals (Cerealia) and orachs (Chenopodiaceae) during the early occupation, detected in the pollen record from Hočevarica (Fig. 1b), which suggests specific farming practices such as crop rotation (Jeraj, 2002). The correlation seems to be very locally restricted since it is not evident from the Bistra pollen record (Fig. 1a), where Chenopodiaceae appeared to be scarcely present and thus are not included in cluster analyses.

In the cluster dendrogram for Hočevarica other high correlations (between 0.8 and 1.0) are shown for alder (Alnus) and hazel (Corylus), for hornbeam (Carpinus) and aquatic plants (Aquatics), and for pine (Pinus) and spruce (Picea) (Fig. 4). High correlations for the first two pairs, Alnus/Corylus and Carpinus/Aquatics, have been also identified in the dendrogram for the Bistra dataset. As already mentioned, the reason for the strong correlation between alder (Alnus) and hazel (Corylus) in the area can be their frequent co-appearance in secondary Holocene phases. Aquatic plants (Aquatics) and hornbeam (Carpinus) both prefer more open and bright areas, which may be the main reason for their co-appearance. The strong correlation between pine (Pinus) and spruce (Picea) may apply to their constant co-presence in coniferous and mixed forest stands around southwestern Ljubljana Moor, which can be detected in the pollen records from the area (Jeraj, 2004). However, their correlation is not so evident from the cluster dendrogram for Bistra (Fig. 2).

Correlations between other plant taxa that appear in the cluster dendrograms for Bistra and Hočevarica are weaker and difficult for ecological interpretation.

4. Conclusions

The LAGRANGE equation discovery method and hierarchical clustering, applied to palaeoecological data from Bistra and Hočevarica, produced considerable results, particularly in terms of discovering correlations among different plant genera/families and their associations growing around southwestern Ljubljana Moor in the past. Some of the found correlations can be well interpreted with the existing knowledge about ecological relations among plant species as well as with available archaeological and archaeobotanical information (e.g. appearance of humans and settlements) from the area. On the other hand, it is interesting that some of the existing relationships among plants are not detected with any of the applied data mining approaches. For example, there is no correlation identified between Abies and Fagus, which have been growing as dominant trees in a silver fir-beech plant community Omphalodo Fagetum around Ljubljana Moor since 7000 years ago. Furthermore, the correlation between Alnus and Corylus that grow in the surroundings as parallel riparian and forest species is detected only with hierarchical clustering analysis. On the contrary, some correlations that are not obvious from pollen diagrams, for example, the correlation between Abies and Cerealia at Bistra, were detected only by applied machine learning techniques.

In conclusion, machine learning techniques applied to pollen datasets from southwestern Ljubljana Moor have successfully detected different patterns in historical compositions of plant communities. However, interpretations of the patterns and correlations discovered using machine learning methods need to be considered with caution because the extrapolation of the results on the region level requires more than two case studies. The knowledge obtained by applied equation discovery and hierarchical clustering methods has

indeed extended present palaeoecological background about formation and succession of plant communities around southwestern Ljubljana Moor in Early and Mid Holocene. In addition, the effects of human activities on Mid Holocene environment during initial occupation were detected.

References

Bell, M., Walker, M.J.C., 1992. Late Quaternary Environmental Change, Physical and Human Perspectives, London, p. 273.

Birks, H.J.B., Birks, H.H., 1980. Quaternary Palaeoecology. University Park Press, Baltimore, p. 289.

Bratko, I., Džeroski, S., Kompare, B., 2003. Analysis of Environmental Data with Machine Learning Methods. Jozef Stefan Institute, Ljubljana, p. 220.

Džeroski, S., Todorovski, L., 1995. Discovering dynamics: from inductive logic programming to machine discovery. J. Intell. Inf. Syst. 4, 89–108.

Džeroski, S., Todorovski, L., Bratko, I., Kompare, B., Križman, V., 1999. Equation discovery with ecological applications. In: Fielding, A.H. (Ed.), Machine Learning Methods for Ecological Applications. Kluwer, Boston, pp. 185–207.

Jeraj, M., 2002. Archaeobotanical evidence for early agriculture at Ljubljansko barje (Ljubljana Moor), central Slovenia. Vegetation Hist. Archaeobotany 11, 277–288.

Jeraj, M., 2004. Archaeobotanical and palaeoecological reconstruction of the southwestern Ljubljana Moor, Slovenia. Doctoral Dissertation, Nova Gorica, p. 136.

Lowe, J.J., Walker, M.J.C., 1997. Reconstructing Quaternary Environments. Addison Wesley Longman, Harlow, Essex, p. 446.

Manilla, H., Toivonen, H., Korhola, A., Olander, H., 1998. Learning, Mining or Modelling? A Case Study from Paleoecology. Discovery Science. Springer-Verlag, pp. 12–24.

Todorovski, L., Cestnik, B., Kline, M., Lavrac, N., Dzeroski, S., 2002. Qualitative clustering of short time series: a case study of firms reputation data. In: Proceedings of the Fifth International Multi-Conference Information Society, vol. A, Jozef Stefan Institute, Ljubljana, Slovenia, pp. 143–146.