Classification of large pollen datasets using neural networks with application to mapping and modelling pollen data

Holmqvist, Björn

2005

Classification of large pollen datasets using neural networks with application to mapping and modelling pollen data

Björn H. Holmqvist

LUNDQUA Report 39
GeoBiosphere Science Centre
Department of Geology, Quaternary Sciences
Lund University

Classification of large pollen datasets using neural networks with application to mapping and modelling pollen data

Björn H. Holmqvist

LUNDQUA Report 39
GeoBiosphere Science Centre
Department of Geology, Quaternary Geology
Lund University

Avhandling för filosofie licentiatexamen

Coden: SE-LUNDBDS/NBGK-05/39+9
ISSN:0281-3076

GeoBiospere Science Centre, Department of Geology,
Quaternary Sciences, Sölvegatan 12, S-223 62 Lund, Sweden
Telephone: +46 46 222 78 80

Lund 2005-03-17

Classification of large pollen datasets using neural networks with application to mapping and modelling pollen data

Björn H. Holmqvist

GeoBiosphere Science Centre, Department of Geology, Quaternary Sciences, Lund University, Sölvegatan 12, SE-223 62 Lund, Sweden
e-mail: bjorn.holmqvist@geol.lu.se

This thesis is based on the three papers listed below, which are presented as appendices I, II and III.

App. I: Bradshaw, R.H.W., Holmqvist, B.H., Cowling, S. & Sykes, M.T. (2000) The effects of climate change on the distribution and management of *Picea abies* in southern Scandinavia. *Canadian Journal of Forest Research* 30, 1992-1998.

App. II: Bradshaw, R.H.W. & Holmqvist, B.H. (1999) Danish forest development during the last 3000 years reconstructed from regional pollen data. *Ecography* 22, 53-62.

App. III: Holmqvist, B.H. Classification of large pollen data sets using unsupervised neural networks. Submitted to *Ecological Modelling*.

Abstract

This thesis concerns the usage of large pollen databases and their application to mapping and modelling past vegetation. Maps of past taxon distributions are generated and classification techniques are used to compile maps of past woodland types. These visualisations of pollen data have applications in forest ecology and in modelling the impacts of climate change. Maps of the distribution limits of *Picea abies* in southern Scandinavia are compared with output from a bioclimatic model to explore distribution-climate relationships during the last 1500 years. Further a classification technique is used to map distributions of Danish forest types over the last 3000 years Classification is done by assigning a sample to a group or a category of similar properties. The categories in this case are woodland types. The classification model is an artificial neural network as trained on an entire database of actual pollen assemblages, resulting in a classification model able to classify pollen samples to a woodland type. This classification model is then used on the grid of interpolated fossil pollen assemblages to produce woodland history maps. Classification methods group the most similar samples, but somewhere a decision has to be made on how many classes or groups to use. I have developed a method for choosing the number of classes that have the highest reproducibility . This is an objective, repeatable method for assessing the optimal number of clusters in a multivariate dataset.

Introduction

The major applications of pollen analysis within scientific research today are in the investigation of continental, regional and local vegetation history and its interaction with climate change and human activities. Within this broad field, there are at least two rapidly developing research areas. Firstly, to understand and quantify the characterisation of pollen source areas (Prentice 1985; Sugita 1994; Sugita et al. 1999). Secondly, the development and application of large pollen databases such as the European Pollen Database (EPD). This is of particular value in data-model comparisons for past climate simulations using General Circulation Models (Prentice et al. 1993), modelling of past biomes (Prentice et al. 1998; Prentice & Webb 1998) and modelling of ecosystem dynamics with dynamic vegetation models such as LPJ-GUESS (Smith et al. 2001). Data model comparisons require well-organised and accessible databases and they also need the data in a form that is directly comparable with model output, for example, tree species distributions and vegetation types rather than raw pollen data. The focus of this thesis is to explore new applications of large pollen databases in the field of data-model comparisons and develop numerical tools that can handle, organise and analyse the very large amounts of pollen data that are now being assembled.

Classification of pollen data

Pollen databases contain information about many taxonomic units at many sites at many points of time. The data can be structured and organised into higher units of taxonomy, space and time to increase their accessibility to modellers. Biomisation is one approach to this organisation where pollen types are assigned to ecophysiological units or plant functional types (PFTs) (Prentice et al. 1996; Prentice & Webb 1998). PFTs are a way of classifying palaeoecological data to make them useful to modellers. They reduce the number of entities considered and provide an ecological rather than phylogenetic basis for treating taxa from different regions in a compatible manner (Prentice & Webb 1998). Pollen types are usually not at the species level but identified to some higher taxonomic category. This can be viewed as disadvantageous by many botanists because of the loss of taxonomic information, but presents no problem to many other data users when the data are re-classified into PFTs and eventually biomes. Once pollen data are assigned to biomes, the changing distribution of biomes in space and time can be mapped. These maps give a record of the biological impacts of past climate change and human activity, which is a major research preoccupation today.

Biomisation has proved to be a very effective way of organising pollen data and extracting maximum possible information that is relevant to climate and vegetation modelling. The process has a few potential weaknesses that leave room for some improvement. Firstly the assignment of pollen taxa to PFTs is essentially a subjective process based upon the experience of the pollen analyst or working group concerned. Quantitative methods are usually more repeatable and scientifically rigorous. Secondly, assignment of pollen taxa to PFTs often uses information about the present distribution of the taxon (from

mapped surface sample datasets) or its host plants from distribution maps (Prentice & Webb 1998). The present vegetation of many parts of the world is heavily modified by human activity, so maps of the present condition may fail to capture much climatic or ecophysiological information.

Biomisation is however only one of several possible approaches to organising large sets of pollen data. Others that have been adopted in the past include assignment of pollen taxa to pre-defined vegetation types (Delcourt et al. 1981), Principal Components Analysis (PCA) (Birks et al. 1975; Huntley & Birks 1983), mapping the 'presumed dominant taxon' (Bennett 1989) and using the phytosociological classification programme TWINSPAN (Hill 1989; Huntley 1990). These are all attempts to organise pollen data into higher units whose changing distributions can be interpreted in climatic and ecological terms. The use of PCA and TWINSPAN are the least subjective of the organisational approaches adopted to date. They are repeatable but have not proved entirely satisfactory for other reasons. The earlier PCA analyses only considered one time slice at a time making comparisons between time periods difficult (Birks et al. 1975). Such ordination techniques can handle larger datasets now so this problem could be avoided. More seriously however, PCA structures the data by placing its first axis to explain maximum variance. In a complex dataset with many merged clusters this may not give the optimal separation of clusters. Consideration of subsequent axes helps to resolve this issue but the results cannot conveniently be displayed on a single map or diagram. The approach I adopt in this thesis treats this problem in a more sensitive manner.

Vegetation types based on TWINSPAN classification appear attractive as due regard can be given to minor taxa that nonetheless have important indicative value. However, the method can yield rather unstable results such as the mapping of a cluster with significant *Picea* content throughout central Scandinavia during the time-period 7000-5000 years BP before the taxon had a significant presence (Huntley 1990).

One of the major aims of this thesis is to quantify the process of classification of pollen data into higher 'ecological' groups and improve upon these earlier efforts. The classification should not be too heavily influenced by the recent landscape so ideally should include data from long periods of time. Such data sets are very large and cannot easily be analysed using conventional classification tools. The classification should reflect as closely as possible any natural clustering within the datasets. It should maximise the use of all data and not be unduly influenced by single pollen types. I use artificial neural networks (Kohonen 1989) as my main analytical tool in this thesis, which appears to have the potential to overcome many of the problems raised within earlier studies.

Neural networks, in this case self organised maps, is essentially a bottom-up approach as they search for underlying structure within the dataset. They identify clusters or nodes and express their relationship in a two-dimensional datamap. I develop methods to identify the optimal cluster number and address a classical issue in vegetation analysis of the relationship between continuities and discontinuities

in vegetation organisation (Gleason 1939; Poore 1956). One scientific question I address is whether it is possible to use quantitative methods to define higher taxonomic groupings that can be interpreted using ecological experience.

Pollen databases

Two sets of pollen data have been used in this study: one covering the whole of Europe taken from the European Pollen Database (EPD) (http://medias.obsmip.fr/paleo/epd/epd_main.html) and one covering Fennoscandia alone, called the NORDMAP pollen database (NM) (Berglund 1991; Birks 1985). NM is a database covering 308 sites from Denmark, Finland, Norway and Sweden, stored at 500 radiocarbon year intervals for the last 10 500 radiocarbon years. It was originally designed for generating vegetation history maps. The selected pollen sites were lakes and bogs exceeding 200 m radius.

The thesis is organised into three papers: the first deals with mapping the distribution limits of a taxon for comparison with a bioclimatic simulation. In the second paper tree pollen data is classified to generate maps of past woodland types in Denmark. The third paper is a further development of state-of-the-art classification techniques for use with international pollen databases.

Summary of papers

Paper 1

Bradshaw, R.H.W., Holmqvist, B.H., Cowling, S. & Sykes, M.T. (2000) The effects of climate change on the distribution and management of *Picea abies* in southern Scandinavia.

***Canadian Journal of Forest Research* 30, 1992-1998.**

In this paper we examine the effects of past climate change on tree distributions at various time-scales, and use this background to study the possible effects of future climate change at a regional scale. We make data-model comparisons to study the role of climate impact in the changing distribution of *Picea abies* in southern Scandinavia during the last 1000 years, a period with reliable proxy records of climate. The distributions of *Picea abies* (L.) Karst. and other European forest trees have continuously changed since the last glaciation. We begin by summarising some of the tree distributions, prior to reliable proxy records of climate. Fossil pollen data held in the NORDMAP database were used to map south Scandinavian tree distributions at 1000-year intervals for the last 8000 years for *Picea*, *Fagus, Tilia* and *Quercus* (Birks 1985; Berglund 1991). The pollen-based tree distributions were interpolated onto maps at thousand year intervals (8000 – present). The maps show that distributions of *Picea*, *Fagus, Tilia* and *Quercus* in Scandinavian have altered significantly and coherently during the last 8000 years. *Tilia* and *Quercus* both extended their ranges northwards to culminate and later on contract southwards. The European ranges of *Picea* and *Fagus* have been constantly expanding during the late Holocene, and they are both 'late' immigrants to Scandinavia.

Static and dynamic bioclimatic simulation models are used to estimate the degree of climatic control operating on the southern Scandinavian range limits of *Picea* during the last 1000

years. The results show that the range limit has begun to track climate change more closely than in the past, and a future projection predicts a rapid northward contraction of the present limit. Though if forest fires become more frequent we anticipate a major role for *Pinus* as occurred earlier in the Holocene (Lindbladh et. al. 2000). Frequent burning gives the fire-adapted *Pinus* a competetive advantage over many deciduous trees, particularly *Fagus*. Contracting ranges track climate change more closely than do expanding ranges that are limited by seed dispersal. The physiological mechanism of the climatic control is unclear. The general exclusion of *Picea* from maritime climates in Europe is probably related to the development and maintenance of winter hardiness that is broken by the repeated freeze-thaw cycles that currently characterise west European winters (Sykes et al. 1996). Recently *Picea* planted beyond its current climatic range limit was seriously damaged during a storm. Planting trees beyond their natural climatic range limits can only be advised in areas that become suitable for colonisation under a changed climate.

Paper 2
Bradshaw, R.H.W. & Holmqvist, B.H. (1999) Danish forest development during the last 3000 years reconstructed from regional pollen data. *Ecography* 22, 53-62.

In this paper we describe tree distribution and forest history of Denmark using quantitative methods that classifies and map pollen data. We use fossil pollen data converted into estimates of tree abundance to map the development of forest types during the last 3000 years. Tree abundance maps of Denmark for selected taxa *Alnus* and *Fagus* show large distribution changes in space and time. We focus our attention on the forested parts of the landscape, even though much of Denmark has had a restricted forest cover for several millennia (Odgaard & Rasmussen 2000). The forest types were clusters in an artificial neural network based on all available European Holocene pollen data. The neural network used was a self organised map which is an unsupervised model that was trained to classify all the data and with the possibility to group any new data presented in the same way. The self organised map generated 32 forest types based on calibrated Holocene pollen assemblages. The types were named according to the dominant taxon or group of taxa. This model applicable to all of Europé, was applied to Danish forest history. Diverse deciduous forest types found 3000 years ago were replaced by less diverse *Fagus*-dominated types over a period of 2000 years. There were significant increases in non-arboreal pollen in southern and eastern Denmark between 3000 and 2500 BP, which corresponds to an opening of the landscape. The association between the increase in non-forest communities and establishment of *Fagus* suggests that anthropogenic activity has accelerated the loss of species-rich deciduous forest with abundant *Alnus, Corylus, Quercus* and *Tilia*. The present day map contained many new combinations of tree species, dominated by *Picea* and *Pinus*. Most present Danish forest types are a direct result of recent silvicultural practice. We conclude that the natural forest composition of Denmark would be deciduous forest today with a significant presence of *Fagus sylvatica*. Recent

forest development has created a break in compositional continuity with the past that is unnatural and has posed problems for forest-dependent biota.

Paper 3
Holmqvist, B.H. Classification of large pollen data sets using unsupervised neural networks. Submitted to *Ecological Modelling*.

In this paper the method created for the application to Danish forest history (App. II) is further refined. The method development can be concluded in three different refinements. Firstly, by finding the optimal number of classes to group pollen data. Secondly to study the reproducibility of several generated classification models and finally to validate the classification by comparing the results with results from randomised datasets. This refined method is also used to analyse large sets of pollen data. Unsupervised neural networks, in this case self organised maps, were trained with pollen data to help describe and visualise past broad-scale vegetation dynamics. Data for this study were obtained from the European pollen database (EPD) and from a specialised Fennoscandian pollen database. Tree pollen data and randomised versions of the data were trained to produce several thousand classification models. Three classification models were selected and studied in detail because they were highly reproducible and well separated from their randomised counterpart. These three models indicated that five was the optimal number of uncalibrated tree pollen assemblage groups in Fennoscandia and nine was the corresponding European number. Transformation of the pollen data to reflect tree composition more closely yielded six optimal groups for the Fennoscandian dataset. A large self organised map was used to analyse the similarities and differences between the three selected models. The two models generated from the Fennoscandian pollen database distinguishes *Picea* presence to a specific boreal class in a Fennoscandian perspective. In contrast, the model generated from the EPD pollen database incorporates this class into a topologically large mixed boreal class. The European model has three much more deciduously pronounced woodland types, dominated by *Corylus*, *Fagus* and *Quercus*. The results were also visualised as vegetation maps for 6000 years ago and present to evaluate the results of the three selected models in a geographical context. This is the first time that large pollen data sets, with all of their multidimensional relationships thoroughly studied, are used to create classification models for the purpose of mapping past vegetation. This method development makes the realisation of palaeovegetation maps possible and they will provide a synthesis of a considerable amount of palynological data. This mapping method will permit evaluation of certain hypothesis about the factors driving vegetational change.

Discussion

For the first time neural network models have successfully been used to map vegetation patterns from pollen data. It is possible that the model with an optimized number of classes achieves robustness, because it is able to identify woodland vegetation communities that show a certain stability through time (App. III). In this case, the neural network classification and mapping is

validated by the many features it has in common to earlier qualitative interpretations of Scandinavian vegetation history (Berglund 1968, 1991; Fries 1965). The approach generates a more objective classification than the groups of pollen types mapped in North America (Jacobson et al. 1987; Janssen & Birks 1994; Williams 2003) and identifies robust clusters more successfully than TWINSPAN (Huntley 1990) because it distinguishes a boreal class with presence of *Picea*. As with most quantitative approaches, this method is repeatable and also being an unsupervised method it is also independent of observer bias. The classifications and data-handling methods presented here are applicable to situations where pollen datasets are to be used in the related disciplines of forest ecology and climate change research. Independent estimates of climate change can be matched against potential vegetation responses. Time periods of particularly rapid change can be identified, and rates of change, and of species range (both expansions and contractions) can be estimated. The age and longevity of different palaeovegetation units can be determined. This contributes to the debate about the nature and timing of vegetation response to climatic change and human impact. In conclusion, the quantitative models described here can generalize large datasets of diverse pollen samples into potential vegetation communities and their history.

Acknowledgements

I thank Richard Bradshaw and Björn Berglund for patient assistance. I also thank Leif Björkman, Mats Holmqvist and Ingmar Kronfeldt for helpful comments.

References

Bennett, K. D. 1989. A provisional map of forest types for the British Isles 5000 years ago. *Journal of Quaternary Science* **4**:141-144.

Berglund, B. E. 1968. Vegetationsutvecklingen i Norden efter istiden. *Sveriges Natur Årsbok* 59:31-52.

Berglund, B. E. 1991. Pollen proxy data from the Nordic countries. Pages 29-36 in B. Frenzel, editor. Evaluation of climate proxy data in relation to the European Holocene. Fischer Verlag, Stuttgart.

Birks, H. J. B. 1985. A pollen-mapping project in Norden for 0-13000 B.P. Nordmap1. Botanisk institutt, Bergen, Norway.

Birks, H. J. B., J. Deacon, and S. Peglar. 1975. Pollen maps for British-Isles 5000 years ago. *Proceedings of the Royal Society of London Series B-Biological Sciences* **189**:87-105.

Delcourt, H. R., D. C. West, and P. A. Delcourt. 1981. Forests of the Southeastern United-States - Quantitative maps for above-ground woody biomass, carbon, and dominance of major tree taxa. *Ecology* **62**:879-887.

Fries, M. 1965. The late-Quaternary vegetation of Sweden. *Acta Phytogeographica Suecica* **50**:269-284.

Gleason, H. A. 1939. The individualistic concept of the plant association. *American Midland Naturalist* **21**:92-110.

Hill, M. O. 1989. Computerized matching of relieves and association tables, with an application to the British national vegetation classification. *Vegetatio* **83**:187-194.

Huntley, B. 1990. European vegetation history: palaeovegetation maps from pollen data - 13 000 yr BP to present. *Journal of Quaternary Science* **5**:103-122.

Huntley, B., and H. J. B. Birks 1983. An atlas of past and present pollen maps for Europe 0-13000 years ago. Cambridge university press, Cambridge

Jacobson, G. L., Jr., T. Webb, III, and E. C. Grimm. 1987. Patterns and rates of vegetation change during the deglaciation of eastern North America. Pages 277-288 in W. F. Ruddiman, and H. E. Wright, Jr., editors. North America and adjacent oceans during the last deglaciation. Geol. Soc. Am., Boulder, CO, United States.

Janssen, C. R., and H. J. B. Birks. 1994. Recurrent groups of pollen types in time. *Review of Palaeobotany and Palynology* **82**:165-173.

Kohonen, T. 1989. Self-organization and associative memory. Springer verlag, Berlin.

Lindbladh, M., R.H.W. Bradshaw, and B. Holmqvist 2000. Pattern and process in south Swedish forests during the last 3000 years sensed at stand and regional scales. *Journal of Ecology* **80**:113-128.

Odgaard, B. V., and P. Rasmussen. 2000. Origin and temporal development of macro-scale vegetation patterns in the cultural landscape of Denmark. *Journal of Ecology* **88**:733-748.

Poore, M. E. D. 1956. The use of phytosociological methods in ecological investigations: IV. General discussion of phytosociological problems. *Journal of Ecology* **44**:28-50.

Prentice, I. C. 1985. Pollen Representation, source area, and basin size - toward a unified theory of pollen analysis. *Quaternary Research* **23**:76-86.

Prentice, I. C., and T. Webb. 1998. BIOME 6000: reconstructing global mid-Holocene vegetation patterns from palaeoecological records. *Journal of Biogeography* **25**:997-1005.

Prentice, I. C., M. T. Sykes, M. Lautenschlager, S. P. Harrison, O. Denissenko, and P. J. Bartlein. 1993. Modelling global vegetation patterns and terrestrial carbon storage at the Last Glacial Maximum. *Global Ecology and Biogeography Letters* **3**:67-76.

Prentice, I. C., J. Guiot, B. Huntley, D. Jolly, and R. Cheddadi. 1996. Reconstructing biomes from palaeoecological data: A general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* **12**:185-194.

Prentice, I. C., S. P. Harrison, D. Jolly, and J. Guiot. 1998. The climate and biomes of Europe at 6000 yr BP: Comparison of model simulations and pollen-based reconstructions. *Quaternary Science Reviews* **17**:659-668.

Smith, B., I. C. Prentice, and M. T. Sykes. 2001. Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. *Global Ecology and Biogeography* **10**:621-637.

Sykes, M.T., I.C. Prentice, and W. Cramer. 1996. A bioclimatic model for the potential distribution of northern European tree species under present and future climates. *Journal of Biogeography* **23**:203-233.

Sugita, S. 1994. Pollen representation of vegetation in Quaternary sediments - theory and method in patchy vegetation. *Journal of Ecology* **82**:881-897.

Sugita, S., M. J. Gaillard, and A. Broström. 1999. Landscape openness and pollen records: a simulation approach. *Holocene* **9**:409-421.

Williams, J. W. 2003. Variations in tree cover in North America since the last glacial maximum. *Global and Planetary Change* **35**:1-23.

Appendix I