REDUNDANCY ANALYSIS AND THE EVOLUTIONARY LEARNING ALGORITHM AS COMPLEMENTARY PROCESSING TOOLS FOR DENDROCHRONOLOGICAL DATA[1].

[1] Paper presented at the IUFRO - All Division 5 Conference - Nancy, France, August 23-28, 1992

H. BEECKMAN* and K. VANDER MIJNSBRUGGE**

* Section of Agriculture and Forestry Economics, Royal Museum for Central Africa, Leuvense steenweg 13, B-3080 Tervuren
** Laboratory of Genetics, University of Ghent, Ledeganckstraat 35, B-9000 Ghent

ABSTRACT

To process dendrochronological data sets, mathematical techniques that can handle complexity are needed. Two methods from the field of numerial ecology are introduced in tree ring analysis : redundancy analysis (an eigenvector method) and the evolutionary learning algorithm (a machine learning tool). Both methods show to be appropriate for a stringent test case. Redundancy analysis explains variance in tree ring data by environmental date revealing main trends. The evolutionary learning algorithm can be applied to look for unexpected strong environmental signals possibly departing from main trends.

Key words : numerical ecology, poplar growth rings, artificial intelligence

RÉSUMÉ

L'analyse des données dendrochronologiques doit tenir compte d'une très grande complexité. Deux méthodes du domaine de l'écologie numérique sont proposées : l'analyse en composantes principales des variables instrumentales ('RDA : redundancy analysis') et l'algorithyme d'apprentissage évolutionaire ('ELA : evolutionary learning algorithm'). On prouve que les méthodes sont toutes deux utiles pour la description des relations entre les cernes et l'environnement : RDA est surtout intéressante pour les tendances majeures, ELA également pour des cas extrèmes avec des forts signaux d'environnement.

ZUSAMMENFASSUNG

Zur Verarbeitung von dendrochronologischen Datenreihen sind mathematische Verfahren notwendig die der Komplexität der Daten Rechnung tragen. Zwei Methoden aus der numerischen Ökologie werden vorgestelt und auf ihre Eignung im gegebenen Kontext getested : der 'redundancy analysis (RDA)' (eine Eigenvektor-methode) und der 'evolutionary learning algorithm' (eine Technik aus der artifiziellen Intelligenz, 'machine learning'). Beide Methoden erscheinen als geeignet. Die 'redundancy analysis' drückt die Varianz der Jahresringdaten in Abhängigkeit von den Umweltdaten als Haupttrends aus. Der 'evolutionary learning algorithm' erlaubt die

Lokalisierung der stärksten Umweltsignale im Datenmaterial.

1. INTRODUCTION

The dendrochronological reality usually appears as a messy laboratory. Puzzling networks of lots of variables and objects, both often of different kinds, are involved (Fritts 1971). Moreover, trees can react in an extremely irregular and unexpected way. A part of this complexity takes shape in the features of the raw unprocessed data matrix (Jeffers 1990) consisting of variables (for example different chronologies) and objects (e.g. the ring widths in distinct years). As far as the variables are concerned, there are often predictands (for example the tree ring widths) and predictors (e.g. environmental data) (table 1).

Table 1 : Schematic representation of a data matrix with predictand and predictor variables.

Predictands	Predictors
e.g. different trees	e.g. different meteorological variables

Predictands as well as predictors sometimes are split up in different categories (table 2). Several chronologies of a species can be filed as one type of predictand variable and chronologies of another species as a second type. As well, different variable types like ring width, vessel area or latewood density can be assorted as different kinds of predictands. Among the predictors, apart from meteorological variables, there could also be variables describing soil characteristics categorised as a second predictor type. Also the objects sometimes are arranged in distinct clusters (table 3) : for example the juvenile and the adult parts of ring chronologies or a partial chronology before and one after forest thinning. Simultaneous clustering within variables and objects provokes the most complex data set (table 4).

Table 2 : Schematic representation of a data matrix with different types of predictands and predictors.

Predictands		Predictors	
e.g. trees belonging to species 1 or variable category 1 (tree ring width)	e.g. trees belonging to species 2 or variable category 2 (vessel area)	e.g. meteorological variables	e.g. soil variables

Table 3 : Schematic representation of a data matrix with different types of objects.

Juvenile objects
Adult objects

Table 4 : Schematic representation of a data matrix with simultaneous variable and object clustering.

variable cluster 1 ↓ object cluster 1 →			variable cluster n ↓
object cluster m →			

A fundamental condition for a mathematical method to be applied in dendrochronology, is that it must be able to handle complexity. Tools to process such data should be looked for among the methods of numerical ecology (Legendre and Legendre 1984), understood as a special category of mathematical ecological methods. Mathematical ecology covers the domain where mathematics is applied in ecology and can be split up in theoretical deductive ecology and quantitative inductive ecology (table 5). Quantitative ecology covers as well ecological statistics as numerical ecology. Ecological statistics applies statistics on ecological problems. Numerical ecology combines statistical methods with other tools which are not dependent on statistical parameters, like mean and variance and are often robust methods, not always bound to rigid stastistical assumptions.

Table 5 : Dendrogram of mathematical ecology (after Legendre and Legendre 1984).

Mathematical ecology

1. Theoretical ecology

2. Quantitative ecology

2.1. Ecological statistics
2.2. Numerical ecology

To process tree ring data, numerous analyses have been proposed. Important numerical ecological tools for dendrochronology are eigenvector analyses (Fritts *et al.* 1971), time series analyses (Visser 1986) and other computer intensive methods like the bootstrap (Guiot 1990).

Because they arrange objects or variables or both in diagrams with variance explaining axes, the eigenvector methods are sometimes called 'ordinations' (Ramensky 1930; Goodall 1954), in human sciences 'scaling'.

Eigenvector analyses are used in tree ring analysis for two main purposes: (1) for calculating a principal or mean chronology (to summarize multidimensional information in diagrams of lower dimension) or (2) for calibration aims (to explain variance among predictand variables by predictors, by means of transfer and response functions).

The eigenvector methods used in ecology can be split up in different categories (table 6). A first distinction is made between direct and indirect gradient analysis (Whittaker

1967). The resulting ordination axes of an indirect gradient analysis are not depending on predictors. The axes could only later, after the main analysis, be interpreted in terms of environmental gradients. Principal component analysis is such an indirect gradient analysis which is frequently applied in dendrochronology since Fritts *et al.* (1971). Data where the complexity consists of predictor and predictand variables can be processed with direct gradient analysis or canonical ordination techniques. Also these kind of methods can be found in tree ring literature and again Fritts *et al.* (1971) can be cited when they used canonical ordination to look for linear combinations between predictors and predictands.

Table 6 : Categories of ordinations (after Ter Braak 1988).

	Indirect gradient analysis	Direct gradient analysis
Linear model	Principal component analysis	Redundancy analysis
Unimodal model	A factor analysis of correspondences	Canonical correspondence analysis

Direct and indirect gradient analysis is only one possible classification of ordination methods. Another division depends on the type of model used for the underlying gradients : analyses exist with linear and unimodal models. A factor analysis of correspondences, proposed for tree ring analysis by Serre (1977), is based upon an unimodal model. A general theoretical model for tree ring data should tend more towards an unimodal form with a lower and an upper limit and an optimum value. Investigating for instance the cambial activity along a long physical temperature gradient, diameter growth is not possible or dormant for as well too low as too high temperatures, but shows an maximum for "optimal" temperatures. However, tree ring data appear frequently to be fairly linear processes, especially when only a short interval of an environmental gradient has been sampled. Also Cook (1990) viewed tree ring series as intrinsically linear processes, possibly after transforming the ring widths to logarithms.

Ordination methods have the power to reveal trends in complex ecological data sets and are therefore in many cases better than calculating chronologies of arithmetic means.

Sometimes however, the description of a principal trend is not of major importance for a dendrochronologist looking for sensitive indicators: sometimes one individual chronology could be a better indicator for an environmental factor than an eigenvector. It is for instance a well known fact that free standing trees on poor soils show stronger meteorological signals.

Furthermore, the complexity of the data set could be too high to be handled by one single eigenvector analysis : sometimes two or more ordinations are needed to process different parts of a chronology, as well as different groupings of variables. To process an extremely complex data set as a whole, no traditional method exists.
However, some methods out of the domain of artificial intelligence are considered as powerful enough to handle such data. With some machine learning tools for instance high quality logical rules can be obtained describing the relation between tree rings and environment. In addition with the same methods it might be possible to focus on chronologies or objects that are extreme or not corresponding to a main trend, but that could reveal an unexpected strong environmental signal. As a consequence these techniques are much promising alternative methods of analysing complex data sets.

Two techniques are being proposed to process tree ring data: redundancy analysis as an eigenvector method and the evolutionary learning algorithm as a machine learning tool. The general hypothesis to verify is "Redundancy analysis and the evolutionary learning algorithm have complementary power to process dendrochronological data."

2. METHODS

2.1. Redundancy analysis

Canonical correlation analysis is a direct gradient analysis processing predictand and predictor variables. While the multivariate model of canonical correlation provides a powerful technique for summarizing and exploring complex relationships between two sets of variables, it is sometimes difficult to interpret and makes a number of assumptions about linearity and the homogeneity of variances and covariances which are seldom true for an actual set of data (Jeffers 1990). A critical constraint results from the specific algorithm of canonical correlation analysis which makes use of parameters estimated through multiple regression. A practical consequence is that the amount of objects must be higher than the total number of variables: the number of rings must be bigger than the number of chronologies, the environmental variables included. This might be a problem for dendrochronological investigations dealing with short chronologies, for example from archaeological or historical objects.

Redundancy analysis is a similar method that doesn't presuppose an upper limit for the ratio between variables and objects and is for this reason to be considered as powerful in a dendrochronological context, but it has not been applied yet.

Redundancy analysis is the canonical form of principal component analysis. The aim is to explain patterns in the data by fitting lines (components). As components, redundancy analysis selects linear combinations of environmental variables that causes the least sum of squares. Redundancy analysis has been developed by Rao (1964) and

has been implemented in the CANOCO-programma, written by Ter Braak (1988).

The output of a redundancy analysis can be decoded by means of biplot diagrams. Rules for constructing and interpreting biplots are given in Ter Braak (1987) : the direction of the vector arrow indicates the direction in which the amount of the corresponding environmental variable increases most; the length of the arrow equals the rate of change in that direction; the angle between arrows of a pair of variables (predictands or predictors) provides an approximation of their pair-wise correlation. The biplot is constructed most easily by drawing seperate plots of objects, predictand and predictor variables, each one with its own scaling. However, within each plot the scale units of the axes must have equal physical lenght. The biplot is obtained by superimposing the plots with the axes aligned and the origins of the coordinate systems coinciding (Ter Braak 1988).

2.2. The evolutionary learning algorithm

The statistical uncertainty of the results of for instance a redundancy analysis can be calculated by Monte Carlo permutation testing. Similar tests have been suggested by Guiot (1990).

Tools that combine Monte Carlo methods with pattern recognition are found in machine learning, a branch of artificial intelligence. Based on information about concrete cases, in the form of descriptive variables about objects, a machine learning programma can assemble so called production rules, rules with IF...THEN...-statements.

The evolutionary learning algorithm, one of the machine learning paradigms, shows some analogy with evolution phenomena (Forsyth 1987). The algorithm starts the analysis by producing at random lots of rules, describing relationships between the variables. A selection of the statistically best rules is made and the programma tries to improve this initial population of rules: components of good quality "parent" rules are recombined (see "cross over" phenomena) or "mutated" to form new "generations" of rules. The rules with the highest quality in terms of predictive power are kept aside, others are killed, which refers to "the survival of the fittest"-phenomena.
The evolutionary learning paradigm has been implemented in the software package Beagle (Forsyth 1987). Beagle is the name of Darwin's ship, but here it stands for Biologic Algorithm Generating Logical Expressions. Beagle consists of six modules. It generates rules that predict a target expression which is a logical formulation of a hypothesis.

2.3. Data set

To provide for a stringent test for both methods, a data set has been chosen consisting of poplar tree rings and meteorological data (Beeckman 1992). The trees grew up in rather favourable conditions without apparent limiting factors. Furthermore the

chronologies are extremely short : only fifteen years. This causes inevitably problems with statistical significance, but also biological problems can't be neglected: a considerable part of the chronologies has been formed by young trees for which the relations with the environment can be totally different from older trees.
The trees were sampled along a spacing gradient : from 3,943 (sample tree a) to 259 (sample tree o) trees per ha (Beeckman 1992). As a consequence the micro environment of the studied trees was obviously different: interference phenomena start later for wide planting distances. Because the chronologies were too short a satisfactory detrending method could not be found. Except for a logarithmic transformation, the data have been standardized by executing principal component or redundancy analysis on a correlation matrix.
The idea behind the choice for such a data set was the assertion that if analyses give here satisfactory results, they should work in more normal dendrochronological context too.

3. RESULTS

3.1. Principal component and redundancy analyses

To quantify the impact of the planting density on the sample, the trees were considered as objects and the years as variables.

A principal component analysis (first four eigenvalues: 0.70, 0.17, 0.06, 0.03) showed that 70 % of the variance can be explained by the first ordination axis. This high value results from a sampling of a rather simple (monoclonal) system with uniform conditions and with one distinct variance explaining factor. Lamarche and Fritts (1971) at the contrary performed a principal component analysis of tree ring data from a vast region in Western United States for the period 1931-1962. In their case only 50 percent of the variance could be explained by the first four ordination axes.

To interpret the ordination axis in terms of a physical gradient, a redundancy analysis was executed with trees as objects, years as predictand and a variable describing the maximum growing area (area potentially available) as predictor. This analysis shows that 70 % of the variance can largely be explained by growing area expressed in m2. The question arises whether some of the remaining variance can be explained by general meteorological variables.

Therefore, in a next run of redundancy analysis, growing area is considered as a covariable to eliminate this main variance explaining factor. This gives an ordination with lower eigenvalues (0.17, 0.08, 0.04, 0.03). A biplot shows that the first eigenvector of this ordination is an indication for low sunshine and low temperatures, the second for high precipitation and the third for high irradiance (figure 1).

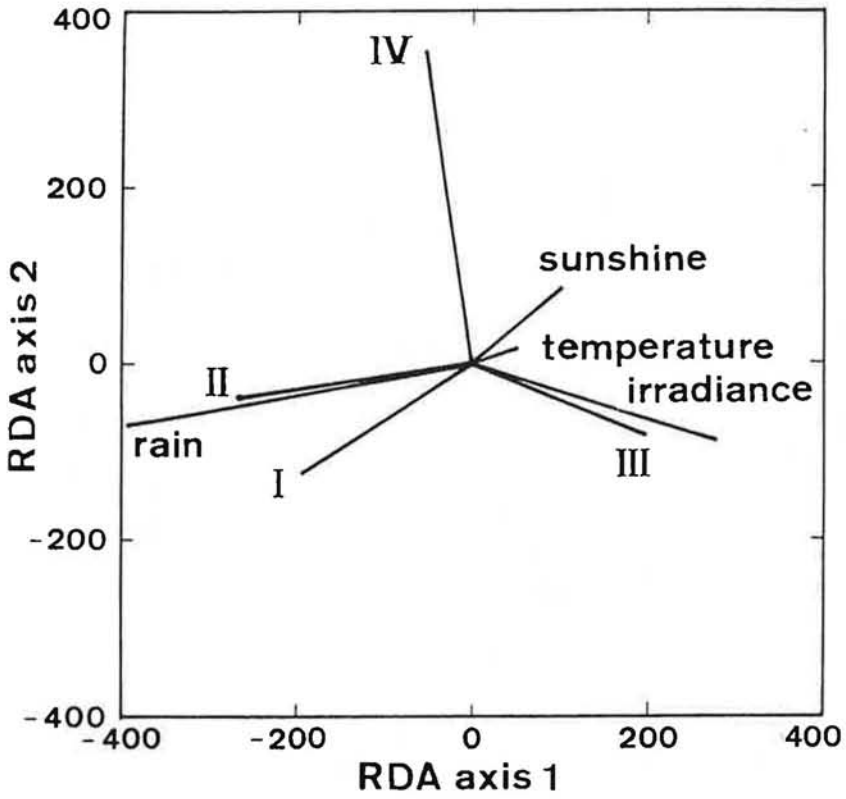

Figure 1 : Biplot diagram of eigenvectors and mteorological variables.

To evaluate the meteorological effects more directly, the trees are considered as predictands, the meteorological variables as predictors and the consecutive years as objects. This is a more conventional point of view in dendrochronology. The results are summarized in a biplot (figure 2). The trees from wider planting patterns tend to have negative correlations with summer temperatures, the others are clearly negatively correlated with amount of summer precipitation.

3.2. The evolutionary learning algorithm

A hypothesis about the effect of mean summer temperatures on tree ring width is formulated as a target expression to be put to a BEAGLE test. As a target expression for instance (temperature > 0) was submitted. Because the tree ring widths are centralized in 0, with this target expression it was possible to look for rules concerning the ring chronologies that express the sensitiveness for summer temperatures higher than the average. The quality of the rules is expressed in contingency tables and the corresponding chi square or related statistics (table 7).

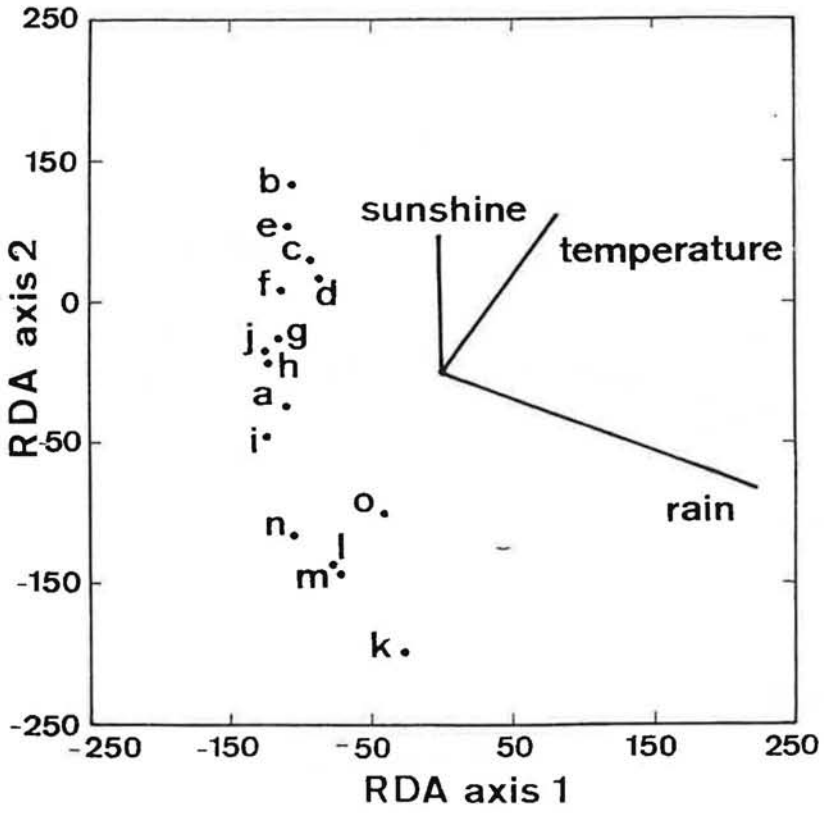

Figure 2 : Biplot diagram of ring width chronologies and meteorological variables.

Table 7 : Contingency table for target expression " temperature > O" and rule " tree 5 > tree 9 ".

		target "temperature > 0"	
		true	false
rule "tree 5 > tree 9"	true	7	0
	false	1	8

With the target expression (temperature > 0), the rule (tree e > tree i) is produced. "tree e" and "tree i" are variables describing the ring widths of the sample trees labelled "e" and "i", which grew up in different spacings.

This can be read as follows. When the summer temperature is higher than the average summer temperature during the period 1974 - 1989 (target expression true), than there is a good chance that the rule will be true too: ring width of tree e (from a high planting density of 1107 trees per ha) will be larger than the ring width of tree i (planting density of 642 trees per ha). The Phi-coefficient (related with the chi square statistic is as high as 87.44.

The combination of chronology e and chronology i in one production rule seems to give a good temperature signal. This result could not be obtained with more traditional processing methods.

4. DISCUSSION

Redundancy analysis has shown to be powerful to explain variability in tree ring patterns of fast growing poplar trees. The data set providing an extreme stringent test set for mathematical methods, makes that the analysis can be considered as particularly suited to handle chronologies which are much longer and originate for instance from alpine or arid regions.

The same can be concluded regarding the evolutionary learning algorithm. This method not only reveals underlying trends in a chronology, but is has also the power to focus on individual chronologies or combinations of chronologies with strong signals not necessarily corresponding to general trends. The method also includes hypothesis formulation, which gives more scope for the researcher.

The choice between one method or another depends on the objectives. The two methods have complementary value: concrete hypotheses can be formulated based on knowledge concerning general trends; down- or upweighting of certain variables in an eigenvector analysis is possible with the information from the results of the evolutionary learning algorithm. Because Beagle's aim is to find a few simple rules that predict a dependent variable or combination of variables, it may throw up some unexpected relationships.

The two methods seem to be powerful enough to be suggested to process dendrochronological data, especially when high degrees of complexity are to be tackled.

5. REFERENCES

Beeckman, H., 1992. Redundancy analysis of tree rings and meteorological data in a Nelder design poplar plantation. In : Bartholin, T.S.; B.E. Berglund; D. Eckstein; F.H. Schweingruber and O. Eggertsson (eds.). Proceedings of the International Dendrochronological Symposium, Tree Rings and Environment, Ystad, South Sweden, 3-9 september 1990. Lundqua Report 34, 22-26.

Cook, E.R., 1990. A conceptual linear aggregate model for tree rings. In : Cook, E.R. and L.A. Kairiukstis, L.A. (eds.). Methods of dendrochronology. Kluwer Academic Publishers, Dordrecht, 98-103.

Forsyth, R., 1987. PC/BEAGLE : user guide. Pathway Research Ltd., Nottingham, 55 p.

Fritts, H.C.; R.J. Blasing; B.P. Hayden and J.E. Kutzbach, 1971. Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate. *J. Appl. Meteor.* 10, 845-864.

Ftitts, H.C., 1971. Dendroclimatology and dendroecology. *Quaternary Res.* 1, 419-449.

Goodall, D.W., 1954. Objective methods for the classification of vegetation, III. An essay in the use of factor analysis. *Australian Journal of Botany* 1, 39-63.

Guiot, J., 1990. Methods of calibration. In : Cook, E.R. and L.A. Kairiukstis (eds.). Methods of dendrochronology. Kluwer Academic Publishers, Dordrecht, 165-177.

Jeffers, J., 1990. Introduction to multivariate analysis. Course on Ecological Statistics. Statistics Group, University of Kent, Canterbury, 1-3.

Lamarche, V.C. and H.C. Fritts, 1970. Anomaly patterns of climate over the Western United States, 1700-1930, derived from principal component analysis of tree-ring data. *Mon. Wea, Rev.* 99, 138-142.

Legendre L. and P. Legendre, 1984. Ecologie numérique. Masson collection d'écologie 12, Tome 1, 260 p., Tome 2, 335 p.

Montgomery, D.C. and E.A. Peck, 1992. Introduction to linear regression analysis. Wiley, New York, 504 p.

Ramensky, L.G., 1930. Zur Methodik der vergleichenden Bearbeitung und Ordnung von Pflanzenlisten und anderen Objekten, die durch mehrere, verschiedenartig wirkende Faktoren bestimmt werden. *Beitrage zur Biologie der Pflanzen* 18, 269-304.

Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. *Sankhya* 26, 329-358.

Ter Braak, C.J.F., 1987. Ordination In : Jongman, R.H., C.J.F. Ter Braak and O.F.R. Van Tongeren (eds). Data analysis in community and landscape ecology. Pudoc, Wageningen, 91-173.

Ter Braak, C.J.F., 1988. CANOCO - A Fortran program for canonical community ordination by partial detrended canonical correspondence analysis (Version 2.1.). Agricultural Mathematics Group, Wageningen, 95 p.

Visser, H., 1986. Analysis of tree-ring data using the Kalman filter technique. International Association of wood Anatomists Bulletin, New Series 7, 289-297.

Whittaker, R.H., 1967. Gradient analysis of vegetation. *Biological Review* 49, 207-264.