O. Brinkkemper

Indirect Correspondence Analysis and Botanical Macroremains: a case study

Correspondence Analysis is one of the multivariate statistical analyses used to explore large data sets. Application to a data set of 106 samples and 246 taxa of waterlogged botanical macroremains from Iron Age and Roman Period settlements on Voorne-Putten (the Netherlands) yielded interesting results. The main distinction between the sites corresponded to their location either in the eastern or in the western part of the area studied. This confirms the results obtained by Cluster Analysis, an alternative multivariate technique. The relation between the samples and the taxa is far more clearly expressed in Correspondence Analysis. Both salinity and moisture appear to be important factors. There is a clear correlation between the occurrence of wild plants indicating salinity and the crop plant barley. This agrees well with experiments revealing the salt tolerance of this crop. Differences in crop plants are also closely related to the dating of the sites. There is a remarkable reduction in the number of crops cultivated during the Roman Period, which can be seen as an indication of arable specialisation.

1. Introduction

One of the previous issues of *Analecta Praehistorica Leidensia*, my thesis (Brinkkemper 1992), dealt with botanical remains from Iron Age and Roman Period settlement sites. The study area comprised the present-day islands Voorne and Putten, situated to the south of the Meuse estuary in the Netherlands (see fig. 1).

An important part of the research concerned botanical macroremains. A total of 107 different samples from eleven sites were analyzed, partly by the present author, partly by W.J. Kuijper. 106 of these samples yielded 246 different taxa preserved in waterlogged, uncarbonized conditions. One sample only revealed carbonized remains. Preservation by waterlogging is usually limited to situations below the water table. In the wet Dutch coastal sites, this means that below ground remains of posts and settlement waste are often preserved in waterlogged conditions.

The possibilities of analyzing the large data set which resulted from the analyses of botanical macroremains "by hand" are severely limited. Our attention is focused on remarkable taxa in the data set, such as crop plants and rare wild plants. The conclusions that can be reached have a fairly limited stretch and are of a haphazard nature. The computer can offer us the possibility of reducing the complexity of the data by means of multivariate analyses. As Lange (1988, 37) stated,

"the tracing of recurrent combinations (synonyms are: associated groups, correlations, covariations, regularities, [...]) in complex data sets is the realm of the Multivariate Analyses of classification and ordination."

One branch of multivariate analyses was used in my thesis, viz. Cluster Analysis. This technique can be characterized as a hierarchical classification technique, where the dendrograms resulting from Cluster Analysis show the relation of the different samples or taxa to each other. Samples or taxa within a given cluster are more similar to each other than to samples or taxa in other clusters.

The data set of waterlogged remains used in my thesis did produce good clusters of the separate sites. The first separation was between the sites in the eastern and in the western part of the study area. Analyses of the taxa yielded clusters which could hardly ever be interpreted satisfactorily from an actuo-ecological point of view. Whether this means that the past vegetation types on Voorne-Putten are not comparable to present ones or whether the data set is of limited value cannot be assessed.

In the introduction to the various multivariate analyses in my thesis, I concluded that the use of Correspondence Analysis would provide a means to explore the robustness of the results obtained by Cluster Analysis. Correspondence Analysis belongs to the second branch of multivariate analyses, being an ordination technique. Ordination techniques search for the largest variation within the data set. The results are generally presented in so-called "biplots", which represent a two-dimensional reduction of the multi-dimensional data set. The first axis accounts for the largest variation within the data, the second axis for the largest remaining variation, and so for the third and higher axes. The distance between the points on the graph is a measure of the degree of similarity or difference. Points close together indicate samples similar

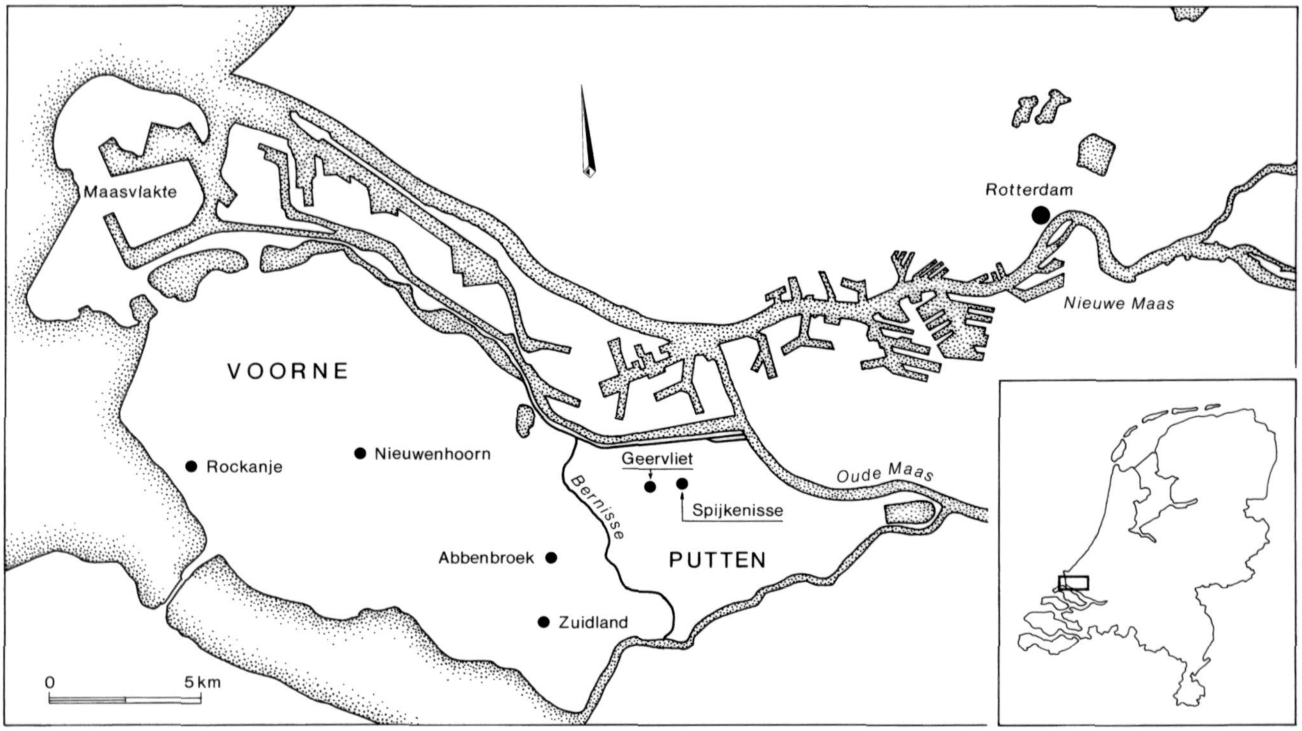

Figure 1. Location of Voorne-Putten in the Netherlands.

in species composition or species occurring in similar samples.

The most commonly applied ordination technique is Principal Components Analysis. A large drawback for its use with palaeo-ethnobotanical data sets is the requirement of a normal distribution of the data. The large number of zero scores in our data sets conflicts with this requirement (compare Jones 1991, 69). Correspondence Analysis does not require such a normal distribution. Furthermore, as Kent and Coker (1992, 203) state, Principal Components Analysis is now widely acknowledged as having serious limitations as a method for the ordination of floristic data. An arch- or horseshoe-shaped distortion in the biplot as a result of this method is one of these limitations.

Correspondence Analysis plots the samples and the taxa against the same axes. This implies that a grouping of samples in a biplot can be interpreted directly in terms of species composition. As Lange (1988, 37) observed, an advantage of Correspondence Analysis and Principal Components Analysis over Cluster Analysis is that both continuous (serial) as well as discontinuous (clustered) patterning may be observed. According to him, Cluster Analysis will produce discrete groups, even when these groups are not present in the data, while ordination would reveal the continuity of the data (see also Sneath/Sokal 1973, 252; Van Tongeren 1987, 174). However, with many clusters, an ordination may give no simple low-dimensional result (Sneath/Sokal 1973, 252).

A considerable advantage of Correspondence Analysis is the possibility to add so-called "environmental variables" to the data analyzed, giving a Canonical Correspondence Analysis. These are extrinsic variables which, in an archaeological context, are either temporal or spatial (*cf.* Jones 1991, 70). Environmental variables used here are dates and contexts of the samples and locations of the sites. Thus, the roles of these factors in any separation of sites and/or samples in Correspondence Analysis can easily be identified. In the default option in Canonical Correspondence Analysis, the axes are based on environmental variables exclusively (see for instance Gaillard *et al.* 1992). This "direct analysis" (*cf.* Kent/Coker 1992, 162) is of great use when determining the effect of these variables on the different species in recent vegetations. In the present study, however, the environmental variables were made passive, *i.e.* they did not contribute to the axes. We can, therefore, speak of an "indirect" analysis, in which the environmental variables are correlated after the variation in the data has been described. According to Kent and Coker, indirect methods can be used in situations where the underlying environmental gradients are unknown or unclear.

Table 1. Archaeological dating and Dutch national coordinates of the sites on Voorne-Putten (after Döbken *et al.* 1992; Van Trierum *et al.* 1988).

Site	Dating	X-coordinate	Y-coordinate
Spijkenisse 17-30	Early Iron Age	80.03	429.86
Spijkenisse 17-34	Middle Iron Age	80.22	429.68
Spijkenisse 17-35	Early/Middle Iron Age	80.275	430.240
Geervliet 17-55	Middle Iron Age	79.300	429.714
Abbenbroek 17-22	Late Iron Age	75.52	427.65
Zuidland 16-15	Late Iron Age	74.33	425.850
Zuidland 17-27	Late Iron Age	75.810	425.250
Rockanje 08-52	Late Iron Age	63.818	432.045
Nieuwenhoorn 09-89	Roman Period	69.660	431.930
Rockanje II	Roman Period	64.44	431.84

2. Methods

The software used for Correspondence Analysis was the CANOCO 3.12 package of the Faculty of Spatial Sciences, University of Amsterdam. In the following analyses, the data were scaled symmetrically, samples and taxa were weighed equally. A ln- or ^{e}log-transformation of the data was used to reduce the influence of taxa occurring in large quantities, such as *Juncus* seeds. Downweighting of rare species appeared to have no influence on the resulting biplot. As some rare species were important in the interpretation of the biplots, they were not downweighted. Two data sets will be discussed below. One set includes all waterlogged macroremains (246 taxa), found in 106 samples. In this data set are 3843 occurrences, which amounts to 15.2%. The remaining 85% of the data set are zero-scores. The second set concerns all remains of crop plants, both waterlogged and carbonized, which occurred with 25 taxa in 69 samples.

The environmental variables considered in this publication are dates, contexts and locations of the sites. The location is expressed in X-coordinates (easting) and Y-coordinates (northing) in accordance with the Dutch national (R.D.) coordinates. These parameters are given in table 1. The contexts and raw data on the macroremains themselves are to be found in tables 10-20 of my thesis (Brinkkemper 1992).

In the following biplots, the diagrams of the samples plus the environmental variables and the diagrams of the taxa are presented separately for reasons of clarity. The axis are identical and the plots can be overlayed.

3. Results and discussion

3.1 WATERLOGGED BOTANICAL MACROREMAINS

The first biplots presented here concern the data set of waterlogged macroremains. The initial Correspondence Analysis included all samples. The resulting biplot showed a dense clustering of all but seven samples. The two samples from Rotterdam-Hartelkanaal, which both yielded very few taxa, were outliers along the second axis. The four samples from the natural subsoils, consisting of raised bogs, in Rockanje 08-52 and Nieuwenhoorn, had high values along the first axis. Similarly high values along the first axis had the sample of goat dung from Nieuwenhoorn, which contained virtually nothing but remains of *Myrica gale*. The low number of taxa for which Correspondence Analysis is sensitive (Ter Braak 1987, 110) causes the extreme position of the samples from Rotterdam-Hartelkanaal and the goat dung, while the low number of taxa and the non-anthropogenic context will be important in the natural subsoil samples. As this is of limited relevance in the interpretation of the data, it was decided to omit these samples in a second analysis. The resulting biplot for the remaining 99 samples is presented in figure 2a.

The eigenvalues of the axes are a measure for the part of the total variation explained (*cf.* Kent/Coker 1992, 187). The eigenvalues of the first and second axes are 0.36 and 0.26 respectively. Ter Braak (1987, 102) stated that values over 0.5 often denote a good separation of the species along the axis. Considering the very large data set, the values obtained here are satisfactory. The sum of all unconstrained eigenvalues is 3.713, so the first two axes account for 9.6% and 7.1% of the total variation respectively.

The separate sites have been indicated by different symbols in figure 2a. The negative scores on the first axis (the left-hand part) exclusively concern sites located in the eastern part of the study area, whereas the sites on western Voorne have positive scores. There is not a single sample that occurs in the "wrong" group. For clarity of the picture, a few samples have been omitted, but these were located in the vicinity of samples from the same site and in the "correct" part of the biplot. A measure of the variation explained by the first and second axes is the squared residual distance of the samples from the plane represented by the axes. Gaillard *et al.* (1992, 13) found a good fit for their modern samples, with low distances (0-10) and a bad

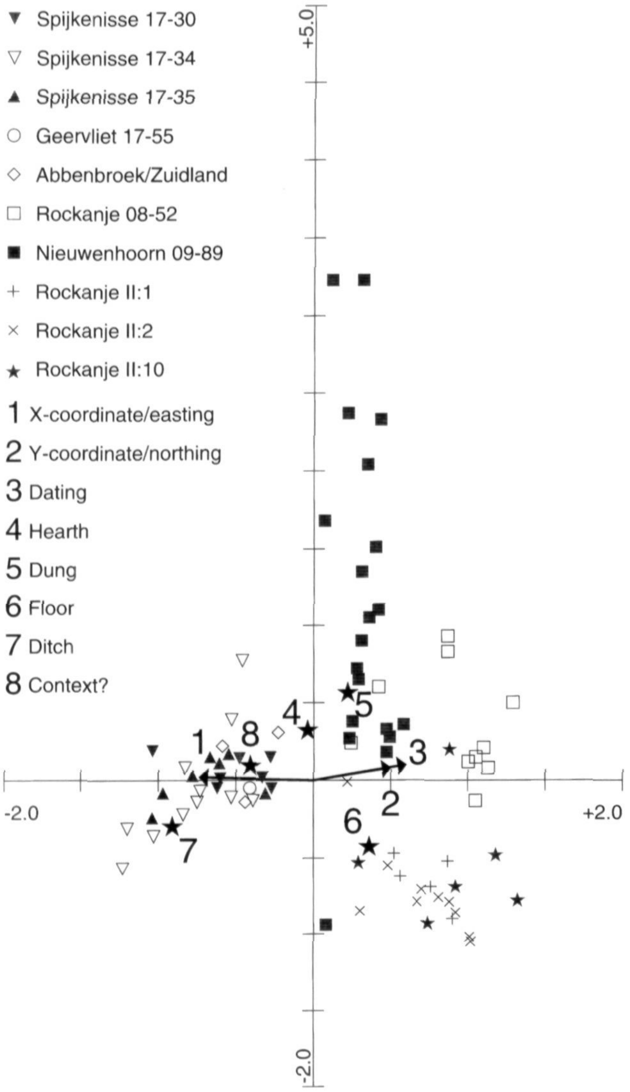

Figure 2a. Correspondence Analysis biplot of samples and passive environmental variables on the basis of waterlogged macroremains.

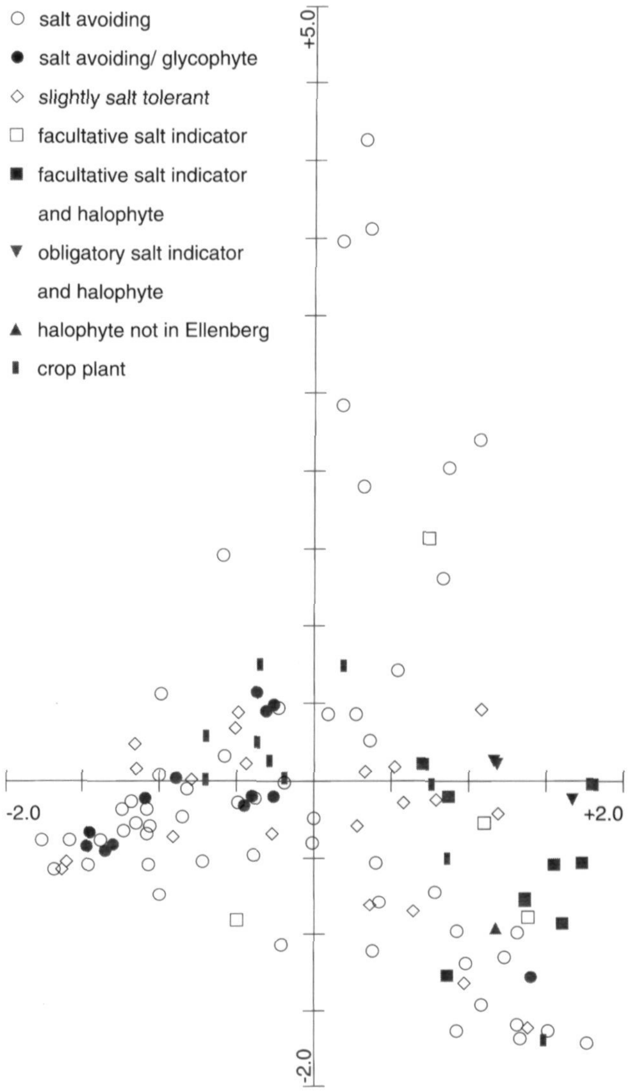

Figure 2b. Correspondence Analysis biplot of species replaced by their salinity indicator value according to Ellenberg 1979.

fit for several fossil spectra (distances >737). The values for the samples from Voorne-Putten range between 0 and 38.3, only sample 10-1-5 from Rockanje II has a value of 645.6. This indicates that only the variation in this last sample is not well explained in the biplots.

The environmental factor X-coordinate (easting) is strongly associated with the first axis, which supports the observed importance of the location of the sites. The factors Y-coordinate (northing) and dating are also mainly directed along the first axis, and have considerable vector-lenghts. This indicates that these are of noticeable influence as well. The fact that the westerly sites are also on average younger than those in the eastern part of the area explains the significance of the dating of the sites. The importance of the Y-coordinate is probably mainly due to the northeast-southwest orientation of the series of sites in the eastern part of the study area, which results in a partial dependence of both variables.

The nominal variables for contexts, indicated by stars instead of vectors, are notable as well. The samples from ditches are grouped in the lower left part of the biplot, those from floor layers on the lower right. The dung and hearth contexts score positive values along the second axis. The separation of the different context types can be seen as an

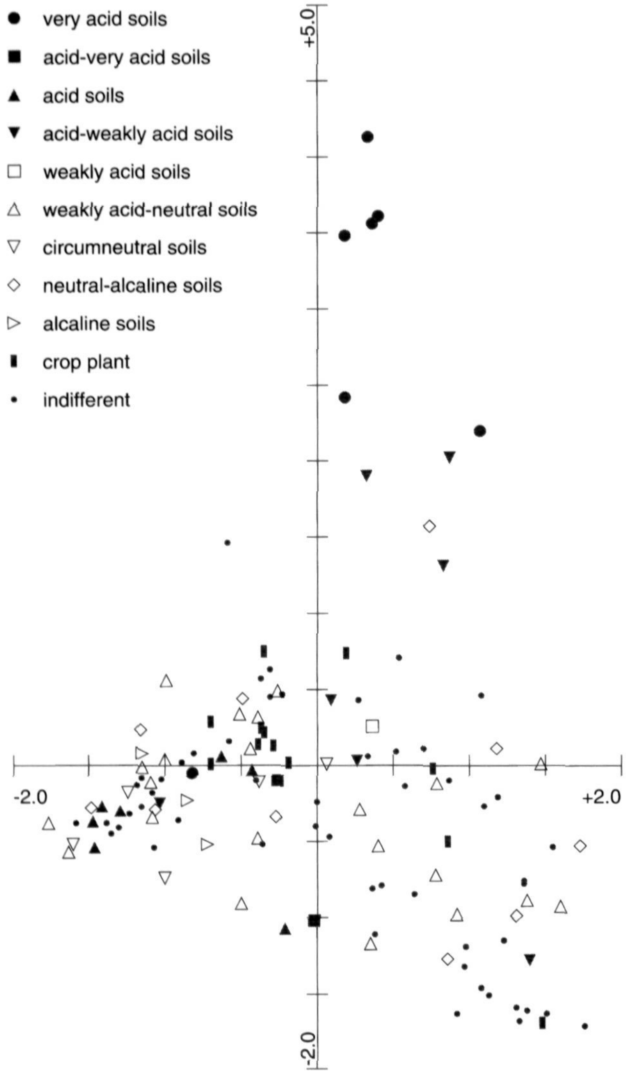

Figure 2c. Correspondence Analysis biplot of species replaced by their acidity indicator value according to Ellenberg 1979.

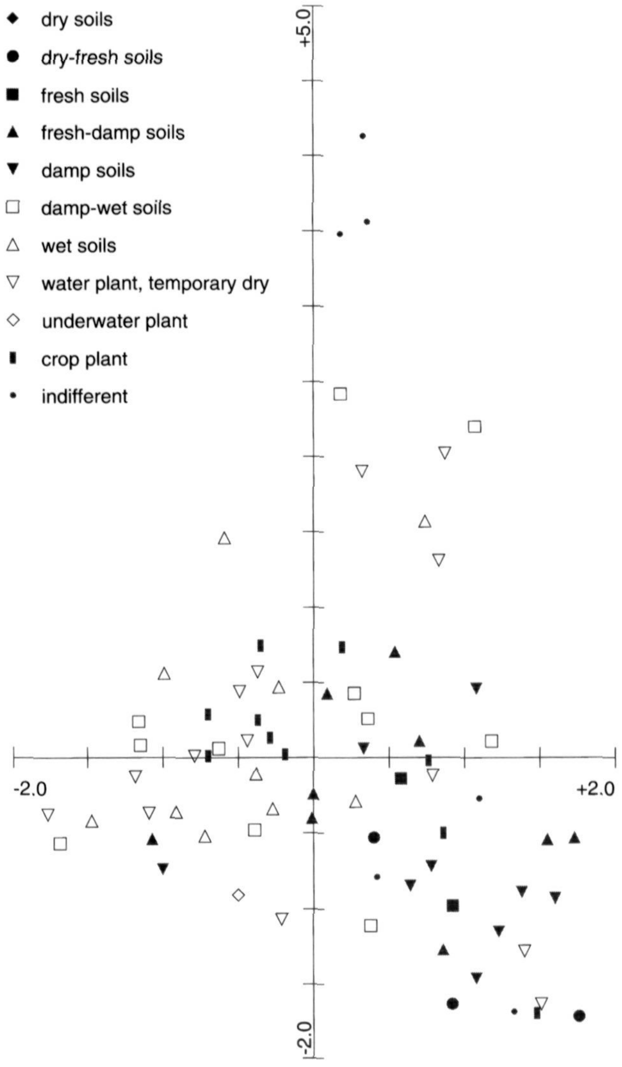

Figure 2d. Correspondence Analysis biplot of species replaced by their moisture indicator value according to Ellenberg 1979.

indication of the necessity to sample as many different contexts on a site as possible to cover the variation in sample contents.

The multiple regression correlation of species and environmental data is 0.97 for the first axis and 0.53 for the second. This indicates that especially the variables along the first axis account for the greatest variation in the species composition (compare Ter Braak 1987, 140).

Corresponding plots for the species are presented in figure 2b-d. A plot containing all taxa names is either unreadable due to overlap or uninformative due to omission of many names. Therefore, the taxa have been grouped according to several criteria. These criteria have been drawn from the study of Ellenberg (1979), who drew up tables classifying species according to their occurrence in relation to abiotic factors. These factors are moisture, nitrogen content and acidity of the soil, salinity, openness of the vegetation and temperature and continentality of the species' distribution. Temperature and continentality are not of relevance here in view of the very small geographical variation between the sites.

In interpreting the results of the Cluster Analyses, salinity was found to be the key factor for the separation of the different sites. Other abiotic factors seemed more or less equally distributed over the sites.

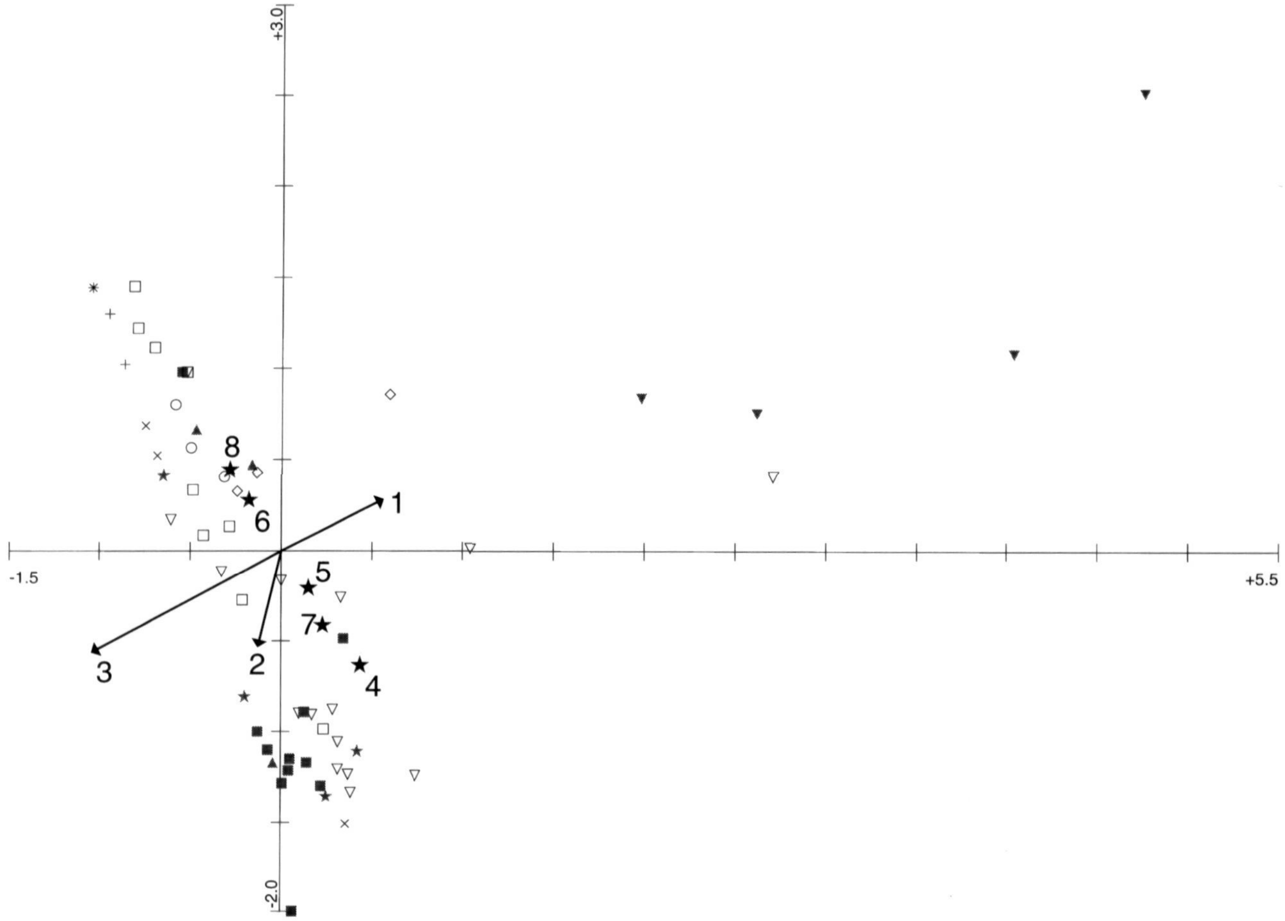

Figure 3a. Correspondence Analysis biplot of samples and passive environmental variables on the basis of all crop plant remains. For legend see fig. 2a.

In figure 2b, the species have been grouped according to their tolerance for salt according to Ellenberg. Besides, taxa which Behre (1985) selected as characteristic for salt (halophytes) or fresh conditions (glycophytes), have been given corresponding black symbols. It should be noted that only identifications to species level can be used here, as genera were not included in Ellenberg's study. Owing to the differences in the ecology of species within most genera, the use of genera is often impossible. Only the higher taxa which Behre included in his study, viz. *Spergularia maritima/salina* or *Rhinanthus* cf *minor*, have been included in figure 2b.

The species biplot shows that nearly all facultative and obligatory halophytes have positive scores on the first axis, whereas the glycophytes occur on both sides of the biplot. The separation of Behre's halophytes and glycophytes is even more pronounced as all halophytes have positive scores and only one glycophyte does. This glycophyte is *Ranunculus flammula*, which is, remarkably enough, limited to samples from Rockanje II.

Interestingly, most crop plants, indicated by black boxes, have negative scores on the first axis as well. The three crop plant remains with distinct positive scores all concern barley remains. Barley is a crop with a relatively high salt tolerance, as has become clear from cultivation of prehistoric crops in saltmarsh conditions (see Körber-Grohne 1967, 230; Van Zeist *et al.* 1977). The association with indicators for saline environments shows the importance of the environment for the agricultural possibilities in the past. In the coast-near sites around Rockanje, the potential of crop plants that could be cultivated successfully was limited in comparison with the sites in the freshwater environment around the Bernisse. Especially the common occurrence of emmer in these latter sites is noteworthy.

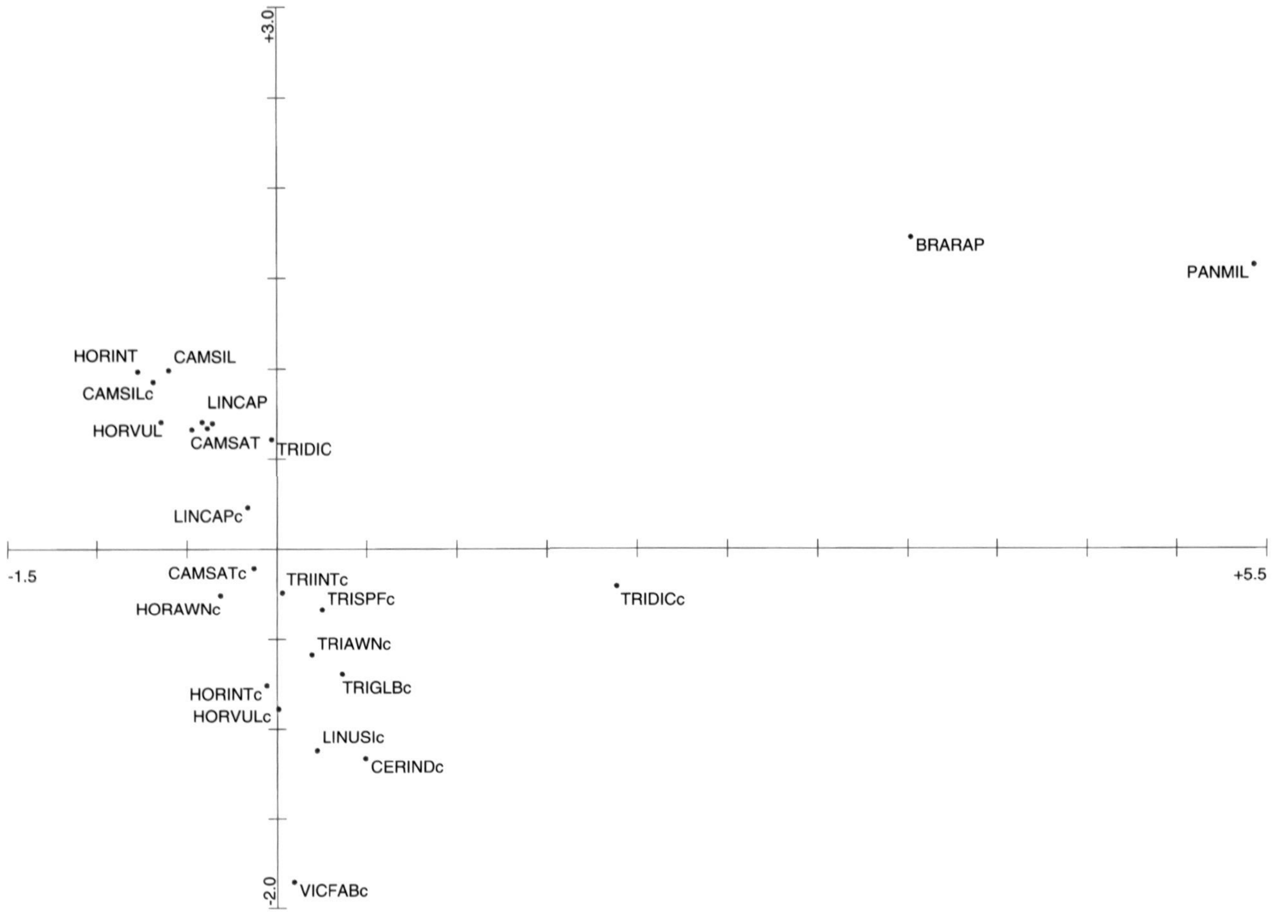

Figure 3b. Correspondence Analysis biplot of taxa on the basis of all crop plant remains.

BRARAP	=	*Brassica rapa*
CAMSAT	=	*Camelina sativa*
CAMSIL	=	*Camelina sativa* silicles
CERIND	=	Cerealia indet.
HORINT	=	*Hordeum vulgare* internodes
HORVUL	=	*Hordeum vulgare*
LINCAP	=	*Linum usitatissimum* capsules
LINUSI	=	*Linum usitatissimum*
PANMIL	=	*Panicum miliaceum*
TRIAWN	=	*Triticum* spec. awn fragments
TRIDIC	=	*Triticum dicoccum*
TRIGLB	=	*Triticum dicoccum* glume bases
TRIINT	=	*Triticum* spec. internodes
TRISPF	=	*Triticum dicoccum* spikelet forks
VICFAB	=	*Vicia faba* var. *minor*
c	=	carbonized

In figure 2c, Ellenberg's acidity values have been used. Clearly, the taxa with high positive scores along the second axis are plants from very acid soils. It concerns *Erica* and *Calluna* remains, which dominate in several samples from Nieuwenhoorn. These samples are located on corresponding parts of the biplot in figure 2a. Other clear trends in the distribution of the different acidity-values are not discernible.

The nitrogen values show a comparable distribution. The above-mentioned taxa are characteristic of very low nitrogen levels as well. Species of very rich and extremely rich soil conditions occur both on the negative and on the positive side of the first axis.

The moisture values (see fig. 2d) show a clear separation along the first axis. Species with lower moisture preferences are concentrated on the right-hand side of the biplot, where the samples from Rockanje II can be found. This site is located near the dry dune area in the western part of Voorne-Putten. Plants from wet environments mainly show negative scores on the first axis. The distribution of crop plants indicate that salinity might have been a more important factor regulating these crops viability than higher moisture values.

The light values show a very restricted range. Shadow and half shadow plants are virtually lacking in all samples.

Thus, the differences in this factor in the biplot are minimal.

3.2 Crop plant remains

The data matrix for crop plant remains included 69 samples with 25 taxa. Carbonized and uncarbonized remains were treated as separate taxa. The biplot of the samples (see fig. 3a) shows extreme locations for samples from Spijkenisse 17-30 along the first axis. The corresponding plot for the taxa (see fig. 3b) reveals that *Panicum miliaceum* (broomcorn millet) and *Brassica rapa* (rapeseed) are responsible for the deviating character of the samples concerned. The remaining samples form a denser cluster, in which most samples from Nieuwenhoorn have relatively low scores along the second axis. The presence of *Vicia faba* (Celtic bean) and the fact that *Triticum dicoccum* (emmer) and *Camelina sativa* (gold of pleasure) are almost absent, causes the separation in the samples from Nieuwenhoorn. The (passive) effects of the different environmental variables, as indicated in figure 3a, is in accordance with these observations. The samples from Spijkenisse are from the early Iron Age, resulting in a considerable score of the vector for dating along the first axis. The samples from Nieuwenhoorn are from the Roman Period, giving an appreciable vector-length along the second axis. The influence of the location of the sites is not as strong as with the waterlogged remains. Besides, they have a bigger influence along the second axis. The vector for dating is longer, indicating that this variable plays a more important role than the location does. This means that there is a time trend discernible in the occurrence of crop plants. It mainly manifests itself in the decrease in the number of cultivated taxa through time, culminating in the virtual absence of crops other than *Hordeum* (barley) in Roman Rockanje. This is a clear indication of increasing specialisation in the cultivation of crops from the Iron Age to the Roman Period.

The nominal variables for contexts mainly reveal a correlation between carbonized crop plant remains and hearths, which is a rather predictable conclusion.

The conclusions for crop plant remains again support and elaborate the results of Cluster Analysis on crop plant remains.

4 Conclusions

The results produced by Correspondence Analysis in the first instance provide a confirmation of the results from Cluster Analysis. The relation between samples, species and the abiotic information drawn from the species, however, is much more straightforward in Correspondence Analysis. The conclusion that salinity is the key factor, explaining the differences in waterlogged macroremains of the different sites, is confirmed. However, the relation of barley with saline conditions is not expressed in Cluster Analysis. Moisture is another abiotic factor which shows a clear trend in Correspondence Analysis, which remained hidden in Cluster Analysis. Passive inclusion of extrinsic environmental variables substantiated the conclusion that the location of the sites, mainly expressed in proximity to the sea, is of great importance in the differences in waterlogged macroremains. It further demonstrates the need to sample as many different contexts as possible on a site.

The Correspondence Analysis of crop plant remains revealed that the samples from Early Iron Age Spijkenisse 17-30 are different owing to the presence of broomcorn millet and rapeseed. Many samples from Nieuwenhoorn are diverging through the presence of Celtic bean and the near absence of emmer and gold of pleasure. This again supports the conclusions drawn on the basis of Cluster Analysis. The interpretation is further aided by passive inclusion of the environmental variables, where the role of dating apparently exceeds the importance of the locations of the sites.

Acknowledgement

The main stimulus for applying Correspondence Analysis to the data from Voorne-Putten was given by G. Jones (Sheffield), both during the Symposium of the International Work Group for Palaeoethnobotany in Kiel as well as during the Workshop "Statistics in Archaeobotany" in Sheffield. W. Schnitger (Amsterdam) gave valuable support with the conversion of the large data set and the printing of figures 2 and 3. C. Brinkkemper (Amsterdam) produced the hardcopies of these figures on film. C.C. Bakels (Leiden) and R. Cappers (Arnhem) gave valuable comments on previous drafts, K. Fennema made the inevitable corrections of the English text.

references

Behre, K.-E. 1985 Die ursprungliche Vegetation in den deutschen Marschgebieten und deren Veränderung durch prähistorische Besiedlung und Meeresspiegelbewegungen, *Verhandlungen der Gesellschaft für Ökologie* 13, 85-96.

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

Brinkkemper, O. 1992 Wetland farming in the area to the south of the Meuse estuary during the Iron Age and Roman Period. An environmental and palaeo-economic reconstruction, *Analecta Praehistorica Leidensia* 24.

Döbken, A.B. 1992 Archeologisch onderzoek in het Maasmondgebied: archeologische kroniek 1987-1990, *BOOR-balans* 2, 271-313.
A.J. Guiran
M.C. van Trierum

Ellenberg, H. 1979 Zeigerwerte der Gefäßpflanzen Mitteleuropas, 2nd Ed, *Scripta Geobotanica* 9.

Gaillard, M.-J. 1992 Modern pollen/land use relationships as an aid in the reconstruction of past land-uses and cultural landscapes: an example from south Sweden, *Vegetation History and Archaeobotany* 1, 3-17.
H.J.B. Birks
U. Emanuelsson
B.E. Berglund

Jones, G.E.M. 1991 Numerical analysis in archaeobotany. In: W. van Zeist/ K. Wasylikowa/ K.-E. Behre (eds), *Progress in old world palaeoethnobotany*. Balkema. Rotterdam, 63-80.

Kent, M. 1992 *Vegetation description and analysis. A practical approach.* Belhaven Press, London.
P. Coker

Körber-Grohne, U. 1967 *Geobotanische Untersuchungen auf der Feddersen Wierde*. Feddersen Wierde, Band 1.

Lange, A.G. 1988 *Plant remains from a native settlement at the Roman frontier: de Horden near Wijk bij Duurstede*. Thesis Rijksuniversiteit Groningen. (= Nederlandse Oudheden 13).

Sneath, P.H.A. 1973 *Numerical taxonomy*. Freeman. San Francisco.
R.R. Sokal

Tongeren, O.F.R. van 1987 Cluster analysis. In: R.H.G. Jongman/ C.F.J. ter Braak/ O.F.R. van Tongeren (eds), *Data analysis in community and landscape ecology*. Pudoc, Wageningen, 174-212.

Trierum, M.C. van 1988 Archeologisch onderzoek in het Maasmondgebied 1976-1986, *BOOR-balans* I, 11-106.
A.B. Döbken
A.J. Guiran

Zeist, W. van 1977 An agricultural experiment in the unprotected salt marsh, *Palaeohistoria* 18, 111-153.
T.C. van Hoorn
S. Bottema
H. Woldring

O. Brinkkemper
p.a. Instituut voor Prehistorie
P.O.Box 9515
NL 2300 RA Leiden