Original article

Canonical correspondence analysis for forest site classification. A case study*

JC Gégout[1], F Houllier[2]

[1] Unité écosystèmes forestiers et dynamique des paysages;
[2] Unité dynamique des systèmes forestiers (associée à l'Inra), laboratoire de recherches
en sciences forestières, Engref, 14, rue Girardet, 54042 Nancy cedex, France

(Received 3 May 1994; accepted 24 July 1995)

Summary – Canonical correspondence analysis (CCA) is an exploratory statistical method that can be applied to the investigation of vegetation–environment relationships and to forest site classification studies. This paper illustrates with a case study some of its advantages over other widely used methods – ecological profiles and correspondence analysis of species abundance data: i) CCA is a global method adapted to the frequent situation characterized by many species and several ecological variables; ii) it makes it possible to underscore the influence of the ecological gradients (eg, water and nutrient availability) on species distribution while eliminating undesirable side effects (eg, the silvigenetic state of the stands); iii) it helps in selecting the ecological variables that are relevant for site classification; iv) it can be used to define synthetic indexes of the ecological optimum and amplitude of plant species and thus to obtain information on good bioindicator species.

site classification / data analysis / ecological gradient / soil–vegetation relationships

Résumé – Analyse canonique des correspondances et typologie des stations forestières. Une étude de cas. *L'analyse canonique des correspondances (ACC) est une méthode exploratoire d'analyse des données qui peut être appliquée à l'étude des relations entre le milieu et la végétation ou pour élaborer une typologie des stations forestières. Cet article illustre, sur un exemple, quelques avantages de l'ACC sur d'autres méthodes classiques – l'analyse factorielle des correspondances d'un tableau phytosociologique, les profils écologiques : i) l'ACC est une méthode globale adaptée à l'étude des relations entre un grand nombre d'espèces et plusieurs variables écologiques ; ii) elle permet d'analyser l'influence des gradients écologiques (exemple : alimentation en eau et niveau trophique) sur la distribution des espèces tout en éliminant des effets parasites (exemple : degré de maturation des peuplements) ; iii) elle permet de sélectionner les variables écologiques pertinentes en vue de la typologie des stations ; iv) elle fournit des indices synthétiques sur l'optimum et l'amplitude écologiques des espèces, indices qui peuvent ensuite être utilisés pour apprécier leur caractère indicateur.*

typologie des stations / analyse des données / gradients écologiques / relations sol–végétation

*Communication at the meeting of IUFRO, Group S1.02.06 'Site Classification and Evaluation', 19–23 October 1993, Clermont-Ferrand, France

INTRODUCTION

The analysis of the vegetation–environment relationships constitutes the central point of forest-site classification studies, which aim at i) determining the ecological gradients that influence the presence and abundance of plant species, and ii) assessing which species are good site indicators. These studies are often based on either plant ecological profiles (Daget and Godron, 1982) or on correspondence analysis (CA) (Hill, 1974; Brethes, 1989).

The method of ecological profiles is analytical (one profile for each pair of species and of ecological variable), it does not account for the redundancy of the environmental variables, nor provide a global overview of the relationships between the ecological gradients and the vegetation.

CA is a global method that is generally applied to plant presence or abundance data. It is most often completed by hierarchical classification methods which aim at grouping sites and/or species (eg, see Buffet, 1984; Roux, 1985). Its main drawback is that it does not lead to a direct analysis of the ecological gradients (Chessel and Mercier, 1993): for example, the first ordination axes sometimes result from the superposition of environmental variables (eg, soil properties) and of forest structure and dynamics (McCune and Allen, 1985; Becker and Le Goff, 1988; Mercier, 1988). A usual way to cope with this problem is to study a posteriori the correlation of the first ordination axes with some external ecological variables (Prodon and Lebreton, 1981).

After Rao (1964) developed the method for principal component analysis, Ter Braak (1986, 1987) and Chessel et al (1987) proposed a new multivariate method that addressed directly the question of vegetation-environment relationships. Ter Braak termed it 'canonical correspondence analysis' (CCA) while Lebreton et al (1988a, b) prefered to name it 'constrained correspondence analysis' or *analyse factorielle des correspondances sur variables instrumentales.*

The aim of this paper is to illustrate with a simple case that CCA is efficient for i) performing a direct gradient analysis, ii) helping the ecologist in the selection of environmental variables that have a strong influence on the vegetation, and iii) assessing the ecological amplitude of plant species.

MATERIALS AND METHODS

Study area

The Plaine de la Lanterne region is located in northeastern France near Luxeuil. Climatic conditions are homogeneous with an average annual temperature of 9.3 °C and an average annual precipitation of 960 mm.year^{-1}. Geological substrata consist of quaternary siliceous alluvium or fluvioglacial deposits, which are frequently covered by a thin loamy deposit (30 to 70 cm). The topography is therefore characterized by gentle slopes (generally < 10%).

Methods

One hundred and six forest sites were sampled in this region (Gégout, 1992). The presence of plant species and environmental variables such as topography, soil characteristics and stand dynamics were observed at each site. The data analysed here are presented in two tables: i) the phytosociological presence/absence table, P, with n rows (n = 106) and p columns (p = 85: only species present at two or more sites were retained); ii) the ecological table, E, with n rows and q columns: the ith row in E as well as in P corresponds to the same site, each column in E corresponds either to a quantitative variable (eg, pH) or to a category of a qualitative variable (eg, the humus form 'mesomull').

Three environmental variables were selected from a previous study (Gégout and Houllier, 1993) and included in table E: 'pH', 'humus form' with six categories (dysmoder and eumoder, hemimoder and dysmull, oligomull, mesomull, eumull, peaty horizon; see AFES, 1992; Jabiol et al, 1994) and 'hydromorphy', an ordinal variable with five categories (absence of hydromorphy, temporary hydromorphy at > 50 cm, temporary hydromorphy at < 50 cm with chroma > 2 at 20 cm, temporary hydromorphy at < 50 cm with chroma ≤ 2 at 20 cm, permanent hydromorphy near the soil surface).

Data analysis

(The computations were carried out with the package ADE [Chessel and Dolédec, 1993] on an Apple Macintosh.)

Since Benzecri (1973), CA has been widely described (Greenacre, 1984). It operates on a single table, here *P*, and yields orthogonal ordination axes that maximize the projected dispersion of either the sites or the plants, the dispersion being defined with the χ^2 metrics (Saporta, 1990). CA generates a summary of *P* that is not a priori constrained by external environmental variables. The ecological interpretation of the ordination axes requires, therefore, the use of such additional variables, which are either plotted on the factorial graphs or correlated with the coordinates of the sites on the first CA ordination axes.

On the other hand, CCA deals directly with two tables, here *P* and *E*. As shown by Ter Braak (1986, 1987), Chessel et al (1987) and Lebreton et al (1988a), CCA may be viewed: i) as a CA of *P* where the ordination axes are linearly constrained by the environmental variables in *E*; ii) as a discriminant analysis between species; iii) or as a CA applied to *P*, the best linear estimator of *P* based on *E*. As a consequence, CCA yields a summary of *P* which depends directly on the environmental variables: i) the intrinsic quality of this summary, as measured by the dispersion projected on the first ordination axes, is necessarily lower or equal to that of CA; ii) the ordination axes can be directly ecologically interpreted.

A usual way for assessing the quality of CA is to compute, $\lambda_{CA,k}$, the eigenvalue associated to the *k*th ordination axis: $\lambda_{CA,1} \geq \lambda_{CA,2} \geq ... \geq \lambda_{CA,k} \geq \lambda_{CA,k+1} \geq ...$. The same quantities, $\lambda_{CCA,k}$, may be computed for CCA and the inequality still holds: $\lambda_{CCA,1} \geq \lambda_{CCA,2} \geq ...$. One of the differences between CA and CCA with respect to this approach is that the number of ordination axes is Min $(n-1, p-1)$ for CA while it is Min $(n-1, q-r)$ for CCA, with *r* being the number of qualitative variables in *E* (a qualitative variable that has *s* classes gives *s* columns in *E*; here $r = 2$ and $s = 6$ for 'humus form'). Since CA provides the best summary of *P*, the following inequality holds:

$$e_m = \sum_{k=1}^{m} \lambda_{CCAk} \Big/ \sum_{k=1}^{m} \lambda_{CAk} \leq 1 \qquad [1]$$

and, as a special case: $e_1 = \lambda_{CCA,1}/\lambda_{CA,1} \leq 1$. e_1, e_2,... can be considered as empirical indexes that measure the efficiency of the ecological variables used in *E* for predicting the structure of the vegetation.

RESULTS AND DISCUSSION

Analysis of the dispersion

The global results concerning the percentage of dispersion are presented in table I. It is limited to the first two axes since the other CA ordination axes had no clear ecological interpretation and had a much lower projected dispersion ($\lambda_{CA,3} = 0.26$, $\lambda_{CA,4} = 0.22$, $\lambda_{CA,5} = 0.19$...). The results have already been presented elsewhere

Table I. Global results of CA and CCA: first two eigenvalues associated to the analyses and relative efficiency of CCA versus CA.

	CA	*CCA on* (P,E)	*CCA on* (P,E)	*CCA on* (P,E')	*CCA on* (P,E')
	λ CA, k	λ CCA, k	e_k *	λ CCA, k	e_k *
	(%)	*(%)*	*(%)*	*(%)*	*(%)*
Total inertia	4.98				
Axis 1	0.45	0.36	80.7	0.33	73.5
Axis 2	0.37	0.20	67.8	0.20	64.2

* See formula [1].

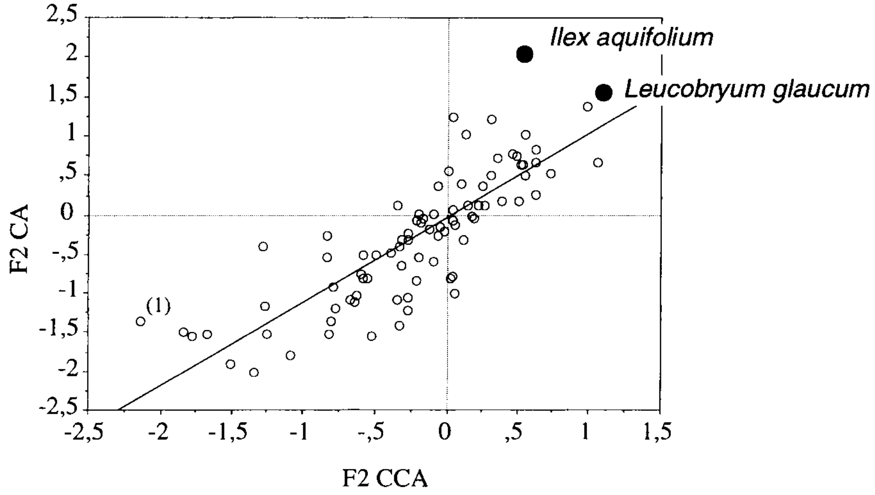

Fig 1. Correlation of the coordinates of the plant species on the second CA axis and the second CCA axis. (1) *Calliergonella cuspidata* is present at only three sites which are characterized by a permanent hydromorphy near the soil surface (two of them with a peaty horizon and the third with an eumull).

(Gégout and Houllier, 1993) and we focus here on the comparison of CA and CCA outputs. CCA is nearly as efficient as CA for predicting the structure of the plant community (e_1 = 0.81 and e_2 = 0.68). The first ordination axis is fairly similar in CA and CCA: the correlation coefficient between species (respectively sites) coordinates is 0.98 (respectively 0.86). This axis accounts for water availability and opposes wet sites to well drained sites. The second ordination axis is more interesting for our methodological purpose here, because its meaning changes from CA to CCA: the correlation coefficient between species (respectively sites) coordinates is 0.82 (respectively 0.57). The CA second axis stems from the superposition of a trophic gradient linked to soil characteristics and a sylvigenetic gradient which opposes pioneer stands to dense mature beech and oak forests, while the CCA second axis accounts only for the trophic gradient and eliminates the effect of the sylvigenetic stages.

This shift of signification of the second ordination axis can be observed by different means. Figure 1 shows that the correlation of the coordinates of the species (on the CCA and CA second axis) is fairly close for those whose presence is strongly influenced by the soil trophic gradient (eg, *Leucobryum glaucum*) but that it is poorer for some species (eg, *Ilex aquifolium*) whose presence is mostly related to the sylvigenetic stage of the stand. Figure 2 illustrates the discriminating role of CCA: humus classes are much better distinguished by CCA than by CA in the plane defined by the first two ordination axes.

For site classification, CCA is shown here to be a more interesting method than the usual CA because it enables us to predict the structure of the plant community from quite simple abiotic environmental gra-

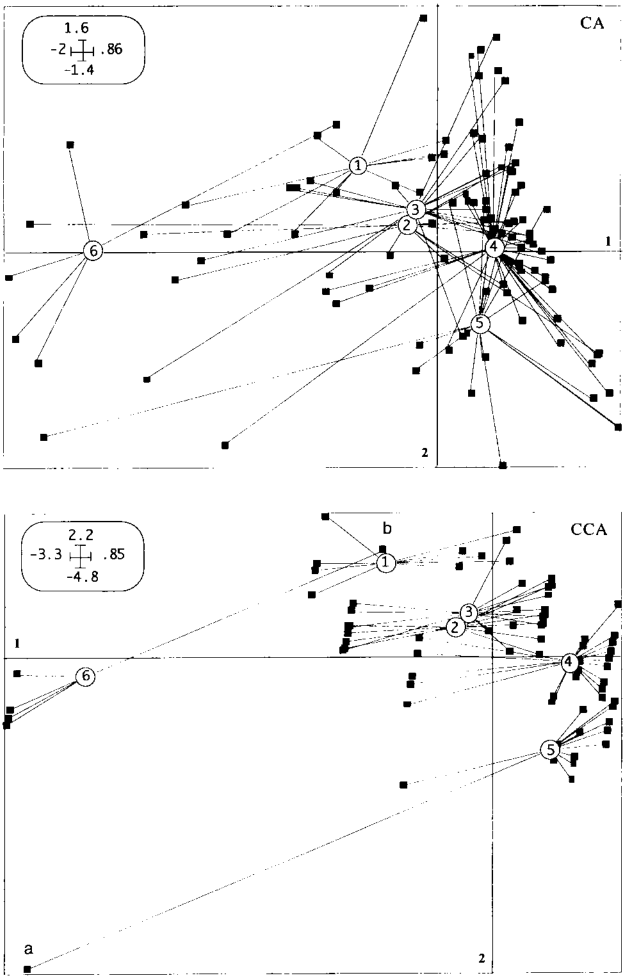

Fig 2. Comparison of the dispersion of humus forms in the first CA and CCA planes. Humus classes are: **1**: dysmoder and eumoder, **2**: hemimoder and dysmull, **3**: oligomull, **4:** mesomull, **5**: eumull, **6**: peaty horizon; a: permanent hydromorphy with eumull; b: temporary hydromophy with peaty first horizon.

dients (water and nutrient availability) and because it eliminates a biotic environmental gradient (the sylvigenetic stage of the stands) that is mainly a consequence of past forest management.

Selection of a set of ecological variables

In order to investigate the pertinence of modifying the description of hydromorphy, CCA was also performed on a second pair of tables *P* (unchanged) and *E*', where hydromorphy was classified in eight categories which account for the intensity of hydromorphy and second horizon chroma (permanent hydromorphy near the soil surface, mottled horizon ≤ 40 cm, 40 cm < mottled horizon < 70 cm, mottled horizon at > 70 cm of depth, some hydromorphic patches without mottled horizon, absence of hydromorphy, chroma at 20 cm ≤ 2 [grey horizon], chroma at 20 cm > 2).

It was not a priori clear whether *E* or *E*' would be best for predicting the structure of the vegetation. The values of e_k in table I

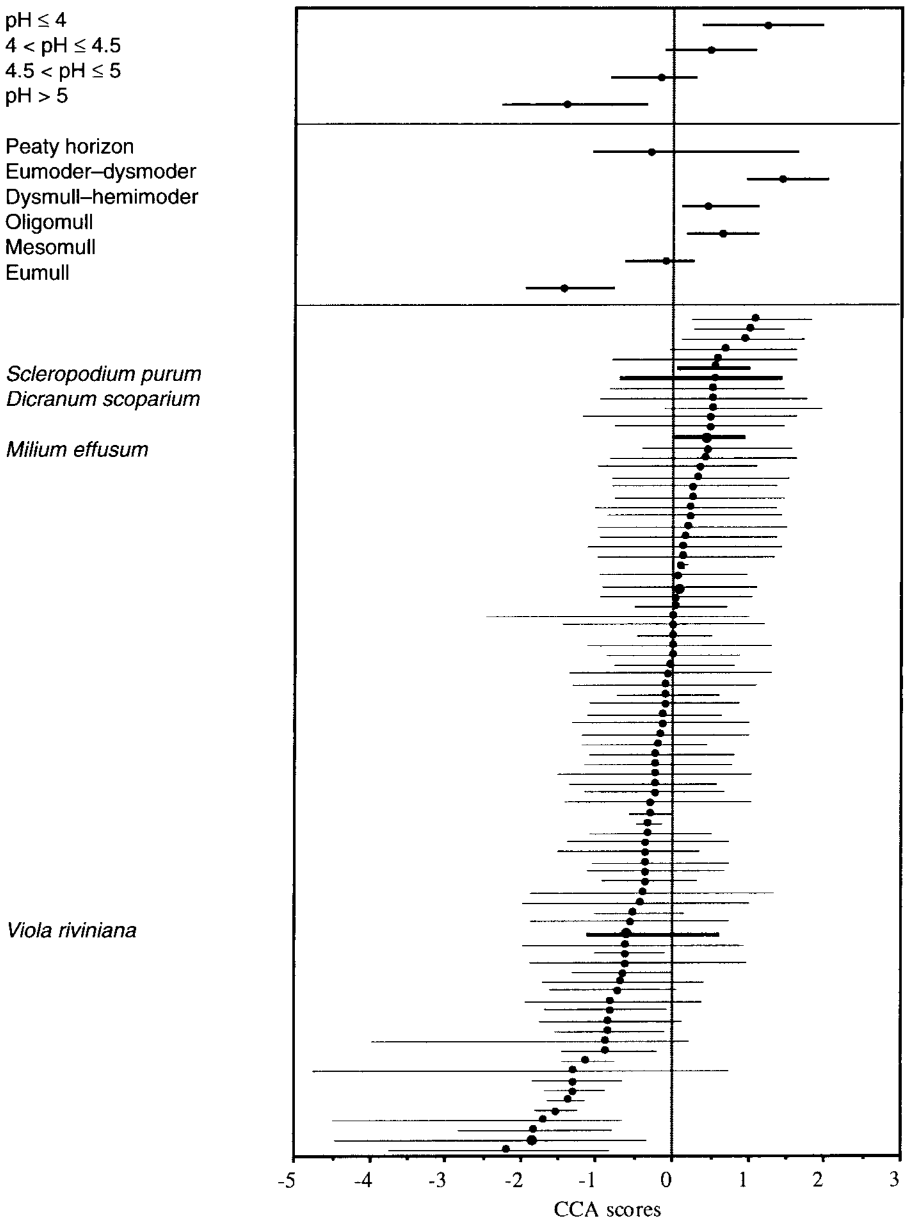

Fig 3. Ecological amplitude of the species along the trophic gradient: average position and.1 and.9 quantiles are figured. Similar information is given for humus and pH classes. Only the species that are cited in the text have been identified.

indicate that E, though simpler, has a better correlation with the vegetation and that it should be preferred to E'. This demonstrates how CCA can be used for detecting which environmental variables are the best predictors of the vegetation. Since there are no statistical tests for comparing e_k from a CCA to another, these ratios should only be used as quantitative indicators of the efficiency of the ecological variables. For example, they can help in investigating whether different categories of the same ecological variables could be merged without altering the discrimination of vegetation types.

Ecological amplitude of plant species

Following Chessel et al (1982) for CA and Lebreton et al (1988a) for CCA, we studied the ecological amplitude of species along the second CCA ordination axis (ie, the trophic gradient) using: i) the coordinates of the species as an index of their ecological optimum; and ii) the coordinates of the sites on the ordination axis to measure their ecological amplitude. This approach is based on the fact that the coordinates of a species are obtained by weighted averaging of the coordinates of the sites where this species is present (Ter Braak, 1986). Precisely, we sorted out the species with respect to their coordinates on ordination axis and computed, for each species, the 1 and 9 quantiles of the coordinates of the sites where it was present (fig 3). This method may be viewed as a multivariate generalization of the analytical technique of ecological profiles (Le Tacon and Timbal, 1973; Daget and Godron, 1982), where the frequency of a species is studied as a function of one environmental variable.

The advantages of the CCA-based approach are manifold. i) As illustrated earlier, the CCA ordination axes are explicitly linked to environmental gradients, while it is not always the case for CA. ii) The

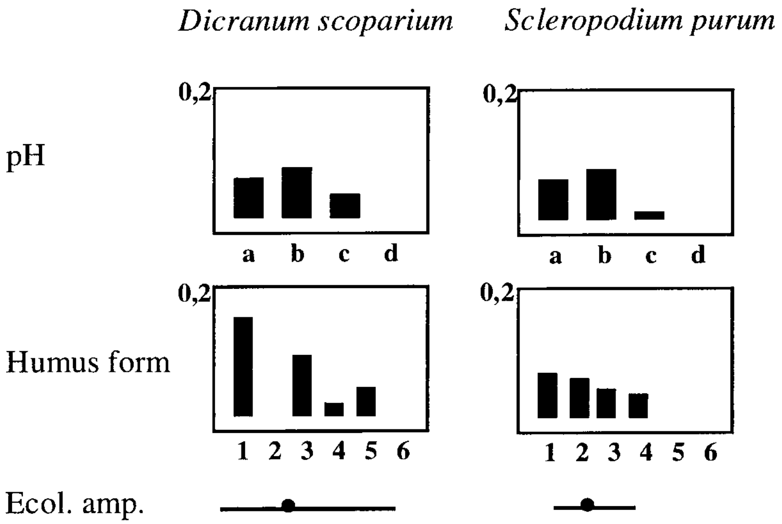

Fig 4. Relative frequency of *Dicranum scoparium and Scleropodium purum* versus pH and humus classes. The ecological amplitude (Ecol amp) along the CCA trophic gradient (see fig 3) is also represented. Humus classes are: **1**) dysmoder and eumoder, **2**) hemimoder and dysmull, **3**) oligomull, **4**) mesomull, **5**) eumull, **6**) peaty horizon. pH classes are: **a**) pH ≤ 4; **b**) 4 < pH ≤ 4.5; **c**) 4.5 < pH ≤ 5; **d**) 5 > pH.

method is global: there are only a few independent ordination axes to study (two in this case). iii) As shown for *Dicranum scoparium* and *Scleropodium purum*, it provides a good description of the real amplitude of the species (fig 4). iv) It can detect nonlinear responses of species to environmental variations. As an illustration, let us take the case of *Milium effusum* and *Viola riviniana* (fig 5). *Milium effusum* is present on dysmull-hemimoder, oligomull and mesomull, which bear approximately the same species (see fig 2); the ecological amplitude of *Milium effusum* is therefore limited. *Viola riviniana* is present mostly on eumull and rarely on oligomull and mesomull. Since these humus classes bear very different vegetation, the estimated ecological amplitude of *Viola riviniana* is broader. The nonlinearity of the vegetation response is clear in figure 3 but not in the ecological profiles given in figure 5.

The utilization of quantiles, instead of standard deviation, provides a nonparametric description of ecological amplitude that can account for asymmetric distributions (eg, *Viola riviniana* in figs 3 and 5). However, since the quantiles of the coordinates are poorly estimated for rare species, the estimated ecological amplitude is highly sensitive to the overall frequency of the various species and thus to the underlying sampling design of the study: this is certainly the major drawback of this method.

CONCLUSION

There are several strategies for classifying forest sites (see Brethes, 1989; Franc and Valadas, 1992). In the context of the phytoecological approach, which is based on the joint study of the structure of the vegetation and of the ecological factors, CCA appears to be a powerful tool that can be complemented by other techniques such as the usual hierarchical classification methods.

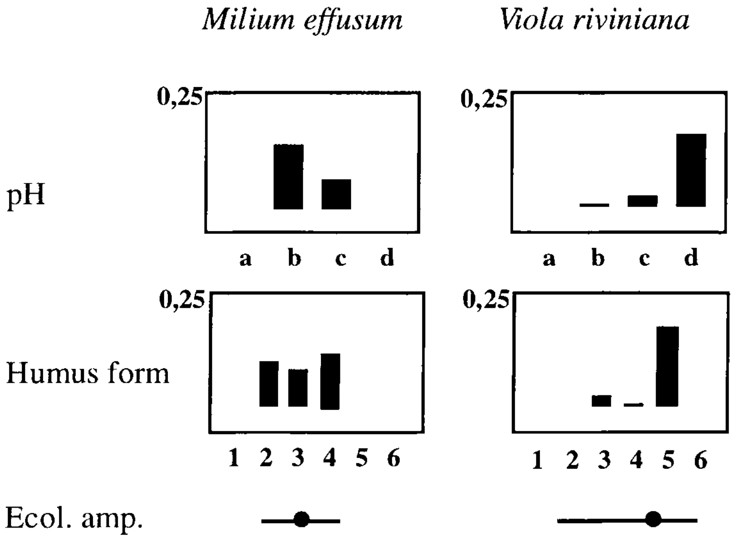

Fig 5. Relative frequency of *Milium effusum* and *Viola riviniana* versus pH and humus classes. The ecological amplitude (Ecol amp)along the CCA trophic gradient (see fig 3) is also represented. Humus classes are: **1**) dysmoder and eumoder, **2**) hemimoder and dysmull, **3**) oligomull, **4**) mesomull, **5**) eumull, **6**) peaty horizon. pH classes are: **a**) pH ≤ 4; **b**) 4 < pH ≤ 4.5; **c**) 4.5 < pH ≤ 5; **d**) 5 > pH.

CCA is therefore a direct method for analysing ecological gradients and, as such, it is a good substitute to the usual two-step approach based on CA for studying the vegetation–environment relationships (Ter Braak, 1986). It may be especially useful for site classification when the environmental abiotic gradients are superposed to other ecological factors that are irrelevant because they are linked to stand physiognomy which is heavily dependent on past forest management.

CCA can be applied as an exploratory method for selecting which ecological factors have the strongest influence on the vegetation and how they should be described (ie, number and nature of the classes for qualitative variables). CCA can also be viewed as a generalization of the one-species versus one-variable approach in order to estimate the relative position and ecological amplitude of the species along environmental gradients.

To a certain extent, CCA is related to the method proposed by Romane (1972) who performed CA on the species versus ecological variables table built by counting the number of times a species is observed for a given class of an environmental variable. Main differences of Romane's approach are that it was symmetric, while CCA is distinctly asymmetric: ecological variables are used to predict vegetation, and sites were not explicitly present, while they appear in CCA.

ACKNOWLEDGMENTS

We are thankful to JL Dupouey, JC Pierrat, S Darracq, JC Rameau, M Buffet and G Vasanthy for their comments on successive versions of the manuscript.

REFERENCES

Association française pour l'étude des sols (AFES) (1992) *Référentiel pédologique : principaux sols d'Europe.* INRA, Paris, 222 p

Becker M, Le Goff N (1988) Diagnostic stationnel et potentiel de production. *Rev For Fr* 40, 29-43

Benzecri JP et al (1973) *L'analyse des données. 2. L'analyse des correspondances.* Dunod, Paris, 620 p

Brethes A (1989) La typologie des stations forestières : recommandations méthodologiques. *Rev For Fr* 41, 7-26

Buffet M (1984) La description du milieu pour l'aménagement des forêts ; application d'un algorithme de classification à la recherche d'une typologie de stations. In: *IUFRO Symposium Aménagement et Gestion (7–11 May 1984, Nancy),* ENGREF, Nancy, France, 31-38

Chessel D, Doledec S (1993) ADE Version 3.6: *Hypercard stacks and QuickBasic Microsoft programme library for the analysis of environmental data.* CNRS URA 1451, université Lyon-I, Lyon, France

Chessel D, Mercier P (1993) Couplage de triplets statistiques et liaisons espèces–environnement. In: *Biométrie et environnement* (JD Lebreton, B Asselain, eds), Masson, Paris, 15-43

Chessel D, Lebreton JD, Prodon R (1982) Mesures symétriques d'amplitude d'habitat et de diversité intraéchantillon dans un tableau espèces–relevés: cas d'un gradient simple. *CR Acad Sci Paris* 295, Série III, 83-88

Chessel D, Lebreton JD, Yoccoz N (1987) Propriétés de l'analyse canonique des correspondances. *Rev Stat App* 35, 55-72

Daget P, Godron M (1982) *Analyse de l'écologie des espèces dans les communautés.* Masson, Paris, 163 p

Franc A, Valadas B (1992) Stations forestières et paysages : les granites du Massif central. *Rev For Fr* 44, 403-416

Gegout JC (1992) *Typologie des stations forestières de la plaine de la Lanterne (Haute-Saône).* ENGREF, Nancy, France, 117 p

Gegout JC, Houllier F (1993) Apports de l'analyse factorielle des correspondances sur variables instrumentales en typologie des stations : illustration sur la plaine de la Lanterne. *Rev For Fr* 45, 539-547

Greenacre MJ (1984) *Theory and Application of Correspondence Analysis.* Academic Press, London, 364 p

Hill MO (1974) Correspondence analysis: a neglected multivariate method. *J R Stat Soc* [C] 23, 340-354

Jabiol B, Brethes A, Brun JJ, Ponge JF, Toutain F (1994) Une classification morphologique et fonctionnelle des formes d'humus. Propositions du Référentiel pédologique 1992. *Rev For Fr* 46, 152-166

Lebreton JD, Chessel D, Prodon R, Yoccoz N (1988a) L'analyse des relations espèces–milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecol* (*Oecol Gen*) 9, 53-67

Lebreton JD, Chessel D, Richardot-Coulet M, Yoccoz N (1988b) L'analyse des relations espèces–milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. *Acta Oecol (Oecol Gen)* 9, 137-151

Le Tacon F, Timbal J (1973) Valeurs indicatrices des principales espèces végétales des hêtraies du nord-est de la France vis-à-vis du type d'humus. *Rev For Fr* 25, 269-282

McCune B, Allen TFH (1985) Will similar forests develop on similar sites? *Can J Bot* 63, 367-376

Mercier P (1988) Approche méthodologique de l'étude des relations entre la structure spatiale du peuplement ligneux et la végétation du sous-bois. *Ann Sci For* 45, 275-290

Prodon R, Lebreton JD (1981) Breeding avifauna of a Mediterranean succession: the olm oak and cork oak series in the eastern Pyrenees. I. Analysis and modelling of the structure gradient. *Oikos* 37, 21-28

Rao CR (1964) The use and interpretation of principal component analysis in applied research. *Sankhya A* 26, 329-359

Romane F (1972) Application à la phytoécologie de quelques méthodes d'analyse multivariable. Discussion sur des exemples pris dans les basses Cévennes et les garrigues occidentales. Thèse de 3^e cycle, université de Montpellier, Montpellier, France, 184 p

Roux M (1985) *Algorithmes de classification*. Masson, Paris, 152 p

Saporta G (1990) *Probalités, analyse des données et statistiques*. Technip, Paris, 493 p

Ter Braak CJF (1986) Canonical correspondence analysis: a new eingenvector technique for multivariate direct gradient analysis. *Ecology* 67, 1167-1179

Ter Braak CJF (1987) The analysis of vegetation–environment relationships by canonical correspondence analysis. *Vegetatio* 69, 69-77