

GENERALISED CANONICAL CORRELATIONS ANALYSIS FOR EXPLAINING MACROFUNGAL SPECIES ASSEMBLAGES

David A. Ratkowsky and Genevieve M. Gates

School of Agricultural Science and School of Plant Science, University of Tasmania, Private Bags 54 and 55, Hobart, Tasmania 7001, Australia. Email: D.Ratkowsky@utas.edu.au and Genevieve.Gates@utas.edu.au

Abstract

Canonical correlations analysis uses a secondary set of multivariate observations to correlate with, and thereby help to explain, differences among a primary set of multivariate observations. In ecology, the variables of the primary data set may be of the "presence/absence" type, or they may be the abundances of individual species in a species list, whereas the secondary set often involves environmental variables. The non-Gaussian distribution of the primary data set precludes the use of traditional multivariate statistical methods, but recent developments in the application of canonical analysis enable similarities between the primary variates to be calculated using measures of similarity in common use in ecology. In this paper, we illustrate how one of these generalisations, the canonical analysis of principal coordinates (CAP), may explain differences in species assemblages of macrofungal survey data obtained from experimental units of differing floristic composition, or how lists of species obtained during different seasons may reflect weather variables. The canonical correlation analyses incorporate non-parametric permutation tests that give exact probability values, as well as ordination diagrams that display the differences in the experimental units that mirror the differences in the mycota.

D. A. Ratkowsky and G. M. Gates (2008). Generalised canonical correlations analysis for explaining macrofungal species assemblages. *Australasian Mycologist* 27 (1): 33-40.

Introduction

A recently developed procedure known as the canonical analysis of principal coordinates (CAP) can visualise differences in macrofungal species assemblage compositions and test for differences among the experimental units using non-parametric permutation tests (Anderson and Willis 2003). The procedure couples principal coordinate analysis (PCoA), an ordination procedure also known as metric multidimensional scaling, with a canonical analysis. The CAP computer program (obtained free of charge from the website of Prof. Marti Jane Anderson of the University of Auckland, <http://www.stat.auckland.ac.nz/~mja>) performs two distinct statistical procedures. The first of these is canonical discriminant

analysis (CDA), in which the experimental units are arranged in predefined groups and the null hypothesis is that the centroids of the groups (i.e. the multidimensional vector of means) do not differ. The permutation test provides an exact, nonparametric P-value for that null hypothesis. An earlier paper in this series (Ratkowsky 2007) illustrated this use of CAP with data from previous macrofungal surveys carried out in the silvicultural trials at the Warra long-term ecological research (LTER) site in southern Tasmania. In essence, these examples tested whether the macrofungal species assemblages differed amongst the groups. The resulting ordination diagrams visualised differences indicated by the permutation tests.

Table 1. Numbers of the large higher plant species in each of the study sites.

Higher plant species	OG	Number of individuals		
		1898	1934	1898/1934
<i>Acacia dealbata</i>	0	0	5	0
<i>Acacia melanoxylon</i>	5	10	18	13
<i>Acacia verticillata</i>	0	0	0	3
<i>Anopterus glandulosus</i>	4	2	33	29
<i>Aristotelia peduncularis</i>	0	0	0	4
<i>Atherosperma moschatum</i>	216	50	3	18
<i>Coprosma quadrifida</i>	5	5	7	28
<i>Cyathodes glauca</i>	0	1	32	29
<i>Dicksonia antarctica</i>	61	43	25	26
<i>Eucalyptus obliqua</i>	2	19	40	39
<i>Eucryphia lucida</i>	8	0	3	78
<i>Gahnia grandis</i>	3	5	23	23
<i>Monotoca glauca</i>	0	2	174	2
<i>Nematolepis squamea</i>	0	0	2	0
<i>Nothofagus cunninghamii</i>	189	86	167	155
<i>Olearia argophylla</i>	0	37	0	17
<i>Phyllocladus aspleniifolius</i>	1	4	22	54
<i>Pimelea drupacea</i>	11	8	7	20
<i>Pittosporum bicolor</i>	0	1	2	1
<i>Pomaderris apetala</i>	0	193	0	832
<i>Tasmania lanceolata</i>	1	0	29	11

The present paper is devoted to the second statistical procedure performed by CAP, namely canonical correlations analysis (CCorA). The objectives of CCorA are quite different from those of CDA, as there are no predefined groups and no hypothesis about group centroids. Instead, there is a second data set of multivariate observations. One possibility is that the first multivariate data set is a list (e.g., presence/absence or abundance) of macrofungal species for each of the experimental units. The experimental units can be a series of plots, divided up into subplots, and there may be a list of macrofungal species for each subplot. The division into subplots may be necessary to provide some form of replication (albeit pseudoreplication in this case) to permit a statistical test to be performed. The second multivariate data set may be made up of the lists of higher plant species or species abundances present in each of the subplots. That is one of the examples illustrated in this paper. A second possibility is that the primary data set may be made up of lists of macrofungal species obtained for each visit to each of the plots, with repeated visits

made over time. The second multivariate data set might be the meteorological data associated with each of the visits, e.g., rainfall and temperature, measured on the days preceding the visits. That is the second example that will be illustrated.

Materials and Methods

Study sites

The data used here to illustrate the CAP procedure were obtained from four 50x50 m plots (each divided into 25 10x10 m subplots) of known fire history chosen along the "Bird Track" at the Warra LTER site in southern Tasmania, Australia. The year of the last wildfire was determined to be (1) ca. 200-300 years ago for the plot designated as Old growth (OG), (2) 1898, (3) 1934, and (4) burnt in both 1898 and 1934, designated as 1898/1934. The four plots were geographically very close, being within 1 km of each other, and were of the same forest type, i.e. wet sclerophyll dominated by *Eucalyptus obliqua*. The stand characteristics of the four plots differ most markedly in the number of *E.*

obliqua and in the understorey, being *Pomaderris apetala* in the 1898/1934 plot and *Monotoca glauca* in the 1934 plot. Table 1 lists the numbers of each of the large higher plant species in each plot.

Visits

Macrofungi, defined as those fungi that produce easily visible fruiting bodies, were recorded during visits made approximately fortnightly. Three of the four plots were visited 30 times, except for 1898/1934, which was visited 29 times, during April 2006 – July 2007, a period which covered two main fruiting seasons. Macrofungal species were recorded on all substrates, including all kinds of wood, both living and dead, soil, litter, moss, dung and ferns. A main focus of the overall project was coarse woody debris (CWD), requiring records to be kept of the fungi that occurred on each piece of CWD, defined in this study to be logs (or stumps) at least 10 cm diameter and at least 1 m long (or tall). Otherwise, each species was recorded once only within each subplot during a visit to a particular plot.

Data matrices for CCorA

First example, first matrix:

CCorA requires two data sets. In the first example to be illustrated, the first data set is the set of macrofungal species obtained from each of the 25 subplots of each of the four plots. A total of 850 macrofungal species was recorded during the 15 months of surveying, so the first data matrix is a 850x100 matrix of macrofungal observations. As each element of this matrix summarises 29 or 30 visits to the given subplot, the entries may contain species abundances or they may be reduced simply to presence or absence (coded as 1 or 0, respectively). We chose the latter option, as some of the species, e.g. the large polypores *Fomes hemitephrus* and *Australoporus tasmanicus*, persist over a long period of time and the same fruiting bodies may be recorded on more than one occasion. Questions such as whether to use abundances or presence/absence in CCorA are not fundamental, as similar results are usually obtained from both methods of coding (the extent of the agreement depending, of course, on the similarity measure employed). It is important to realise, however, that neither the

use of abundances nor presence/absence approximates the assumption of multivariate normality that is required by the standard, classic canonical correlations analysis. In the classic case, both sets of multivariate observations are required to be normally distributed in order that the multiple correlation coefficient produced by the process may be interpreted statistically, and a P-value assigned to it. The modification of CCorA embodied in CAP does not assume multivariate normality. Instead, similarities between the lists of macrofungal species in the subplots may be based upon measures favoured by ecologists, such as the Bray-Curtis measure. The statistical test used by CAP is distribution-free, employing random permutations.

First example, second matrix:

The second data set in the first illustrative example is a 100 x 21 array of stand characteristics, the rows being the 100 subplots of the four plots combined, and the columns being the 21 plant species given in Table 1. The purpose of the CCorA is to see whether the stand characteristics correlate with differences in the macrofungal species lists among the 100 subplots. As this is a correlation analysis, there is no real distinction between the concepts of a "first" and a "second" data set. The two matrices can be interchanged and the results will be the same, there being no dependent variables as there are in multiple regression analysis or its multivariate extension, redundancy analysis (see, for example, Legendre and Legendre 1998, for further information on these procedures).

Second example, first matrix:

The second example, in common with the first example, has a first matrix of 850 macrofungal species recorded during the 15 months of surveying, but this time the lists involve visits rather than subplots. As there were 30 visits to three of the plots, and 29 visits to the remaining plot, the first data matrix is a 850x119 matrix of macrofungal observations. Note that although species were recorded separately for each subplot at each visit to a plot, the data were combined into a single list at the plot level.

Second example, second matrix:

The second data matrix of the second illustrative example is a 119x12 matrix of weather observations, the rows being the 119 visits to the four plots combined, and the columns including the 3 day, 7 day and 14 day rainfall total, the 3 day, 7 day and 14 day average minimum temperature, 3 day, 7 day and 14 day average maximum temperature, and the 3 day, 7 day and 14 day average mean daily temperature.

Data transformation, standardisation, and distance measure:

In both examples, macrofungal species records from each subplot or visit were converted to presence/absence and the distances between the experimental units calculated using Bray-Curtis dissimilarity, without standardisation or transformation.

Results

First example:

The permutation test for the canonical correlations analysis (CCorA) carried out by CAP for the data in the first example gave a P-value of 0.000200 when 4999 permutations were used. Since $0.000200 = 1/(4999 + 1)$, this means that no randomly permuted data set had a more extreme macrofungal species assemblage than that of the original data set. We can therefore conclude that there are real differences among the lists of species in the 100 subplots, and that these differences most likely reflect the stand characteristics. Additional CAP output assists the user in determining differences, such as the correlations of each of the canonical axes with the macrofungal species and the correlations of the canonical axes with each of the higher plant species. However, we will look first at the graphical output of CAP, to see if differences between the information in the 100 lists may be visualised.

Fig. 1 displays the first three canonical axes of the CCorA, plotted in two parts, as Axis 2 vs. Axis 1 (left) and as Axis 3 vs. Axis 1 (right). The 100 points, representing the 100 subplots of the combined four plots, are represented by plotting symbols chosen to differentiate the four plots. There appears to be a clear separation of points based upon plots, with the

twice burnt plot 1898/1934 being the most different. The other three plots have some degree of overlap, suggesting that other factors may be at work. Therefore, the next step is to use stand characteristics when choosing the plotting symbols. Five of the most common plant species with stems ≥ 5 cm diameter in each subplot were plotted on the same graph (Fig. 2). Comparing Fig. 2 with Fig. 1 reveals that because 12 of the subplots from the 1898 plot have *Pomaderris* as their most abundant tree, those 12 subplots appear in Fig. 2 with 24 of the subplots from 1898/1934 characterised by the same species. Furthermore, the 1934 plot has been split mostly into two unequal parts, the larger being the 12 subplots with abundant *Monotoca*, and the smaller having *Nothofagus* predominating, the latter group joined by some subplots originating from Old growth and some from 1898. It remains unclear, however, whether the stand characteristics are a more compelling separator of the differences among the full list of 850 macrofungal species than the plots themselves. This suggests that it may be informative to use both the plots and the stand characteristics simultaneously; this can be done by choosing plotting symbols which identify plot differences and stand characteristic differences in the same graph. This composite approach is shown in Fig. 3, using a dozen different plotting symbols that represent combinations of the plot and the floristic component. This appears to be a more successful way of displaying the differences between the species lists in the 119 sampling units (i.e. the plot visits) than the two previous attempts in Figs 1 and 2.

We now look at other CAP outputs. Table 2 gives the correlation coefficients between each of the 21 stand characteristic species and the canonical axes. If we choose 0.5, somewhat arbitrarily, as a cut-off point for whether a correlation coefficient is sufficiently large, then only the first three of 16 canonical axes have any correlations that exceed 0.5 in absolute value (Table 2). The first canonical axis is most strongly associated with *Pomaderris*, followed by *Phyllocladus*, both with negative signs. *Pomaderris* was present in all 25 subplots of 1898/1934, often in large numbers, and *Phyllocladus* was present in 22 of the 25 subplots of that plot. This helps explain the location of the subplots of 1898/1934 on the

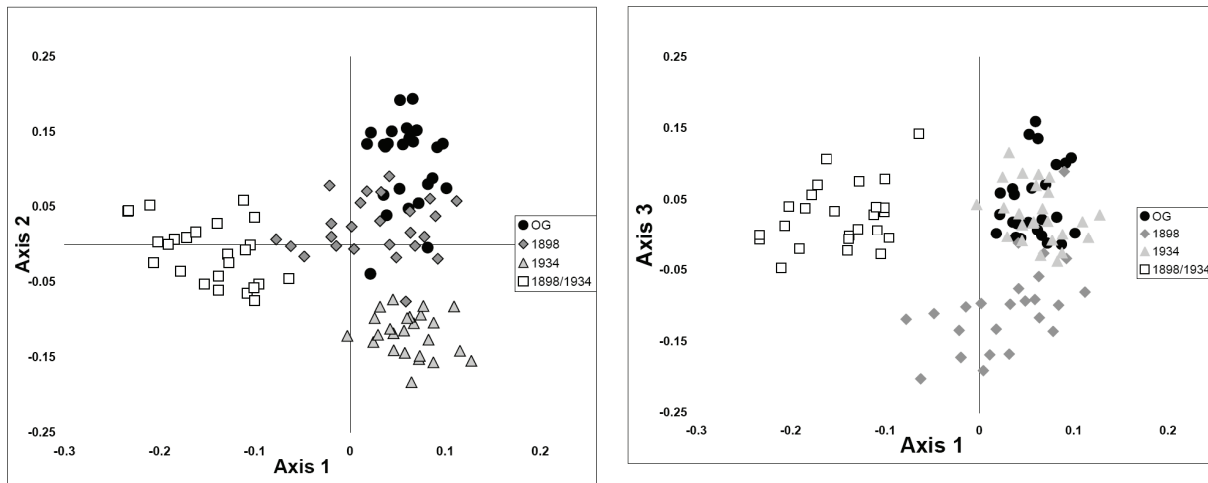


Figure 1. The first three canonical axes of the CAP analysis, displaying the 100 subplots using plotting symbols identifying the four plots to which each subplot belongs. (L): Axis 2 vs. Axis 1, (R) Axis 3 vs. Axis 1.

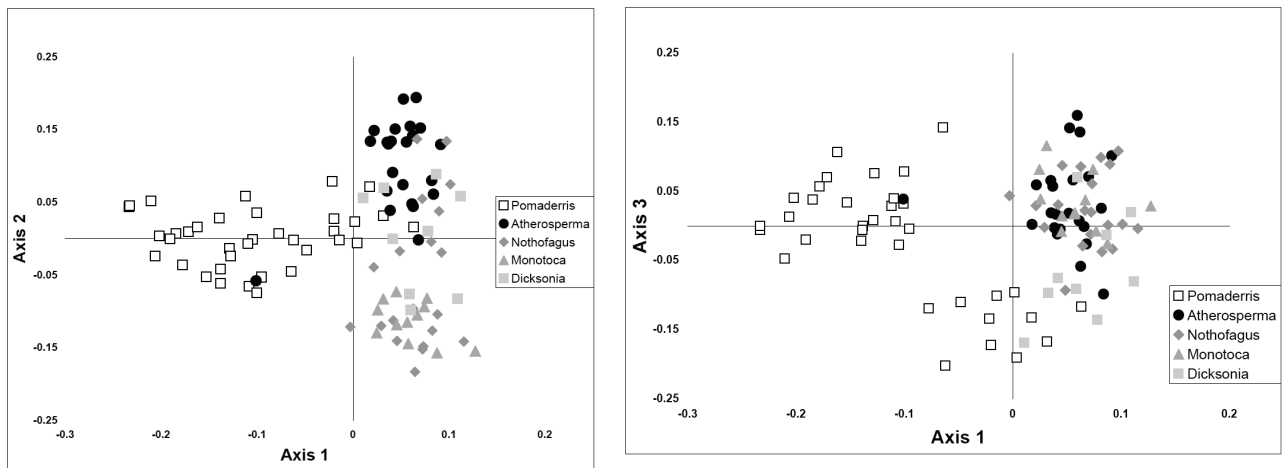


Figure 2. The first three canonical axes of the CAP analysis, displaying the 100 subplots using plotting symbols identifying the most frequently occurring plant species in the subplot. (L): Axis 2 vs. Axis 1, (R) Axis 3 vs. Axis 1.

negative-valued side of Axis 1 of Figs 1 and 3, and also most of the subplots with a major component of *Pomaderris* on the negative-valued side of Axis 1 of Fig. 2. Axis 2 is most strongly positively correlated with *Atherosperma* and most strongly negatively correlated with *Monotoca* and *Gahnia*. *Atherosperma* dominates 21 Old growth subplots, as indicated by the position of the points at the positive side of the scale of Axis 2 of Figs 1-3. Similarly, the points representing the subplots of 1934 on the negative end of

Axis 2 indicate the presence of *Monotoca* and *Gahnia* there. As for Axis 3, the largest correlation is with *Olearia argophylla*, a species present in 15 subplots of 1898, but infrequent in 1898/1934 and absent from other plots. This helps, in part, to explain the location of most of the subplots of 1898 at the negative side of Axis 3 of Figs 1 and 3. One must always bear in mind that the axes are determined by all of the 21 plant species and that although a few species may have fairly

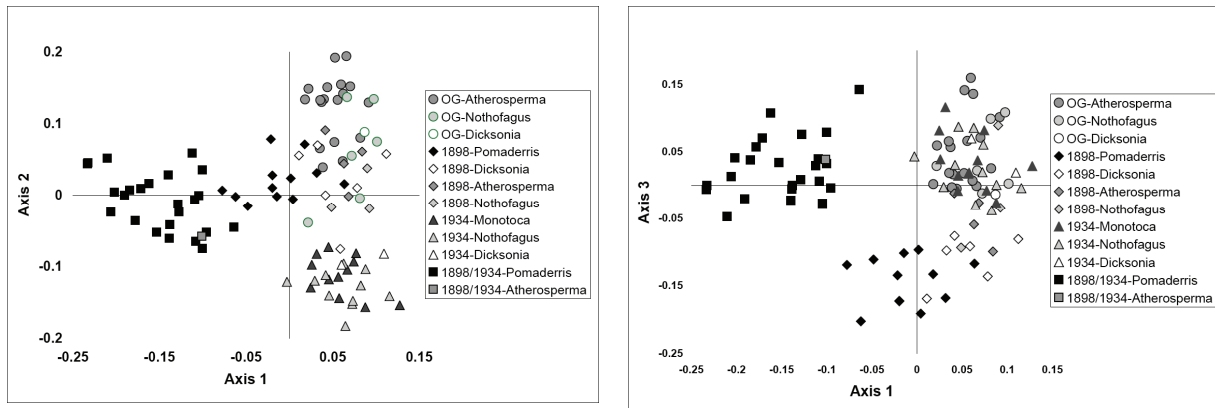


Figure 3. The first three canonical axes of the CAP analysis, displaying the 100 subplots using plotting symbols which are a composite of the plot identifier and the most frequently occurring plant species in the subplot. (L): Axis 2 vs. Axis 1, (R) Axis 3 vs. Axis 1.

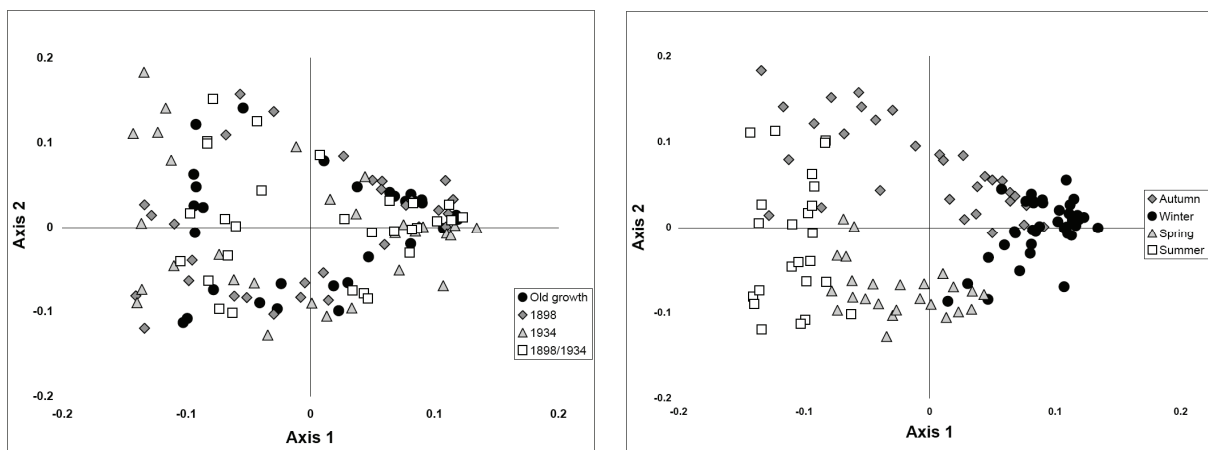


Figure 4. The first two canonical axes of the CAP analysis, relating macrofungal lists to weather data: (L) the 119 subplots are displayed using plotting symbols identifying the four plots, (R) the 119 subplots are displayed using plotting symbols identifying the four seasons.

large correlations, other species also contribute to some extent.

The correlations between the plant species and the canonical axes are one of the two sets of correlation coefficients produced by CAP. The other is the set of correlations between the macrofungal species and the canonical axes. Although these correlations might be considered useful for identifying indicator species, they generally do not prove to be of value. The reason is that the canonical axes are constructed as linear combinations of the set of all 850 macrofungal species in such a manner as to correlate most highly with

another set of linear combinations, namely the set involving the stand characteristics. The fulfilling of this aim has nothing specifically to do with finding good discriminators between, for example, pairs of plots, or among the floristic components found in the set of 100 subplots.

Second example:

The data for the second example involve the effect of weather on the macrofungal species that were recorded at each of the 119 visits. The permutation test (4999 permutations used) indicated real differences between the macrofungal species lists and that the

Table 2. Correlations of canonical axes with each of the stand characteristic species. Large correlations are shown in boldface.

Species	Canonical axis		
	1	2	3
<i>Acacia dealbata</i>	0.09	-0.20	0.13
<i>Acacia melanoxydon</i>	-0.04	-0.29	-0.07
<i>Acacia verticillata</i>	-0.24	-0.06	0.07
<i>Anopterus glandulosus</i>	-0.07	-0.20	0.25
<i>Aristotelia peduncularis</i>	-0.20	-0.05	0.13
<i>Atherosperma moschatum</i>	0.31	0.69	0.28
<i>Coprosma quadrifida</i>	-0.41	-0.05	0.07
<i>Cyathodes glauca</i>	-0.16	-0.30	0.13
<i>Dicksonia antarctica</i>	0.24	0.32	0.07
<i>Eucalyptus obliqua</i>	-0.32	-0.45	0.03
<i>Eucryphia lucida</i>	-0.29	-0.07	0.17
<i>Gahnia grandis</i>	-0.37	-0.62	0.18
<i>Monotoca glauca</i>	0.26	-0.50	0.13
<i>Nematolepis squamea</i>	0.05	-0.13	0.03
<i>Nothofagus cunninghamii</i>	0.12	-0.11	0.26
<i>Olearia argophylla</i>	-0.21	0.06	-0.55
<i>Phyllocladus aspleniifolius</i>	-0.59	-0.30	0.16
<i>Pimelea drupacea</i>	-0.24	0.06	0.13
<i>Pittosporum bicolor</i>	0.03	-0.06	0.19
<i>Pomaderris apetala</i>	-0.88	0.03	-0.07
<i>Tasmannia lanceolata</i>	0.04	-0.40	0.22

Table 3. Correlations of canonical axes with each of the weather variables. Entries greater than 0.5 in absolute value are shown in boldface.

Weather variable	Canonical axis		
	1	2	3
3-day rainfall	0.14	0.04	-0.04
7-day rainfall	0.14	-0.13	0.10
14-day rainfall	0.18	-0.32	0.13
3-day ave. minimum temperature	-0.64	0.19	0.15
7-day ave. minimum temperature	-0.81	-0.13	0.10
14-day ave. minimum temperature	-0.78	0.00	0.13
3-day ave. maximum temperature	-0.69	0.30	0.20
7-day ave. maximum temperature	-0.84	0.04	0.11
14-day ave. maximum temperature	-0.80	0.15	0.15
3-day mean daily temperature	-0.73	0.40	0.13
7-day mean daily temperature	-0.88	0.15	0.05
14-day mean daily temperature	-0.84	0.26	0.09

differences correlate with the rainfall and temperature records on the days prior to the visits ($P=0.0002$). Fig. 4 shows the first two canonical axes of the CAP analysis, graphed using plants and seasons as symbols, respectively. Table 3 gives the correlation

coefficients of each of the canonical axes with each of the weather variables. From Fig. 4, it is clear that any differences between plots are insignificant compared to seasonal weather differences. Rainfall has little or no effect (Table 3); the main separation on Axis 1 is due

to the differences in temperature between summer and winter months. Autumn and spring are spread out between the extremes, reflecting the intermediate temperatures that characterise these seasons.

Discussion

Both examples have successfully applied canonical correlations analysis to a research problem. The first example related the abundances of plant species to the presence/absence lists of fungi, and the second example related weather variables to presence/absence lists. In each example, permutation tests indicated a highly significant relationship between the macrofungal species lists and other variables that may aid in explaining the fungal communities present. These examples prove that CCorA can be useful for understanding the complex interactions of fungi in forest ecosystems. This gives ecologists another tool at their disposal for analysing complex multivariate data sets without having to satisfy the restricted assumption of multivariate normality. Although the canonical correlations analysis option of CAP has not been used as frequently as the canonical discriminant analysis option, it should not be overlooked as a technique for interpreting multivariate ecological data.

REFEREES FOR VOLUME 26

The editorial board thanks the following people for refereeing manuscripts for the journal in 2007: Tina Bell, Mark Brundrett, Richard Ford, Cheryl Grgurinovic, Paul Guy, Pavel Kalac, Lorelei Norvell, David Orlovich, Barbara Paulus, Ceri Pearce, Shaun Pennycook, Morten Strandberg.

Acknowledgements

This study was supported by an Australian Postgraduate Award and grants from Forestry Tasmania, the CRC for Forestry, the Bushfire CRC, the Holsworth Wildlife Research Endowment Fund, and the University of Tasmania Schools of Plant Science and Agricultural Science.

References

- Anderson, M.J. & Willis, T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology* **84**, 511-525.
- Legendre, P. & Legendre, L. (1998). *Numerical Ecology*. Second English Edition. Elsevier, Amsterdam.
- Ratkowsky, D.A. (2007). Visualising macrofungal species assemblage compositions using canonical discriminant analysis. *Australasian Mycologist* **26**, 75-85.

CALL FOR PAPERS

This year we will be publishing a special issue on medical mycology, edited by Assoc. Prof. Wieland Meyer. All members and non-members are invited to submit research papers and reports on any aspect of medical mycology electronically to:

A/Prof. Wieland Meyer, Molecular Mycology Research Laboratory CIDM, ICPMR, Level 3, Room 3114A, Westmead Hospital, Darcy Road Westmead, NSW 2145, Australia. Email: w.meyer@usyd.edu.au

Deadline: June 30, 2008.