Journal of Vegetation Science **23** (2012) 1013–1023

Species presence/absence sometimes represents a plant community as well as species abundances do, or better

J. Bastow Wilson

Keywords
Abundance scales; Analysis stability; Assembly rule; Community structure; Data quality; Environmental correlation; Recording errors; Survey; Vegetation

Received 22 August 2011
Accepted 22 April 2012
Co-ordinating Editor: Janos Podani

Wilson, J.B. (bastow@bastow.ac.nz): Botany Department,University of Otago, Dunedin, PO Box 56, New Zealand

Abstract

Question: Many broad-scale surveys are made, and local communities described, with time spent recording some measure of the abundance of each species. The results are always somewhat different from those obtained with presence/absence records, but which best represents the underlying structure of the community? Two partial answers to this question are suggested and tested here: which analysis correlates best with the habitat, and which gives more stable ordination scores under subsampling?

Methods: Tests were made on ten field data sets, ranging widely in habitat type, spatial extent, spatial grain and measure of abundance. Correlation with the habitat was examined for the four larger-extent data sets with reasonably complete environmental information, using multiple regression of detrended correspondence analysis (DCA) ordination scores on environmental factors. Stability was tested for each data set using random subsets of the quadrats, and measuring stability as correlation between quadrat ordination scores in the subset and those using all quadrats.

Results: Correlation with the habitat for the four data sets, where possible, was closer with presence/absence in most comparisons. Stability was greater with presence/absence in some cases and with abundance in others. Where abundance analyses were more stable, reduction to abundance categories, which are often used in field sampling, resulted in a loss of stability, although in two out of three data sets some advantage of abundance information over presence/absence was retained. Jittering to simulate subjective recording gave no further degradation.

Conclusions: The data sets in which stability was higher in abundance analyses suggest that abundance is of value only in the rather homogeneous vegetation types that tend to occur over short distances, and with high-quality abundance data. From this, and environmental correlations being on the whole better with presence/absence analyses, I conclude that in broader-scale survey work, abundance information is unnecessary and may even be misleading. It seems that the primary assembly rule control on communities is on the presence of species, not their abundance.

Introduction

Broad-scale mapping of vegetation types is increasingly important in ecology (e.g. Bunce et al. 1996; McDonald 1997; Schaminée et al. 2007; Jennings et al. 2009), and countless surveys are made on smaller scales, for example to describe the variation within a conservation area or to address a specific hypothesis. Usually an estimate is made of the abundance of each species in each quadrat.[1] But is this necessary?

Possible measures of abundance include biomass, density, sub-quadrat frequency, cover by point quadrats,

[1]I shall use 'quadrat' throughout, although 'site', 'sample', 'plot' or 'relevé' might be used in an original paper. I use 'abundance' to include any quantitative measure of the amount of a species in a quadrat (Kershaw 1985).

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

repeated cover, cover by line intercept, cover estimated subjectively, and subjective abundance categories (Wilson 2011). All represent trade-offs between speed, objectivity and realistic representation of the community, but all are more time-consuming than recording species presence/absence. For example, sub-quadrat frequency can take about four times as long as recording only presence/absence (Økland 1990). Practical considerations, in large-scale surveys often restrict abundance assessments to a subjective estimate, perhaps on Braun-Blanquet +, 1–5 or Domin +, 1–10 scales, or cover categories. Even this takes time.

In measuring treatment effects in an experiment, biomass or repeated cover is clearly needed. Repeated cover, cover or local frequency is probably needed for determining temporal change. Otherwise, the measure used may depend on the purpose to which the analysis will be put. For example, if indicator species are required for quick survey purposes, an analysis based on presence/absence, and therefore more easily definable in those terms, may be preferable (Dale 1995; Dufrêne & Legendre 1997). But in general for community description, is any measure of abundance necessary? Might presence/absence give a result that represents the community, as it is seen after environmental filtering and assembly rule restrictions, as well as any measure of abundance, or (unthinkably) a better one?

The issue is also theoretical: of whether the control on species occurrence in a community is primarily on whether a species can persist there or whether on the other hand it is easy for a species to enter and persist in small amounts, but the control is on the abundance that the species reaches. This is basically an issue of assembly rules, 'restrictions on the observed patterns of species presence or abundance that are based on the presence or abundance of one or other species or groups of species' (Wilson 1999, 2007; Wilson & Stubbs 2012). In this case, the question is whether the species assemble primarily by presence or by abundance.

Comparisons of presence/absence and abundance analyses

Some authors have performed ordinations on the same data set using both presence/absence and abundance information ('information' is used here in the sense of Orloci 1966, p. 193: 'the information content of a matrix,' and Feoli et al. 2006, p. 100, 'information on the floristic composition of plant communities,' not in the mathematical usage as seen in the Shannon–Weaver index of diversity). For example, Orloci (1966) found that PCA ordinations of a sand dune/slack data set based on presence/absence and frequency data were only slightly different. Similarly, Avena et al. (1981), with Braun-Blanquet-scale data from Italian oak forests, found 'no substantial differences' between presence/absence- and abundance-based canonical variates ordinations (Feoli & Orloci 1979). However, they judged that a classification of the data based on presence/absence predicted species abundances in the groups better than an abundance-based classification predicted species presence. Perhaps this is not surprising, since presence/absence cannot have half-correct predictions. In contrast, Vermeersch et al. (2003) concluded that DCA ordinations of an area of forest in Belgium using only presence/absence information were 'unsatisfactory' in that they did not match traditional phytosociological associations well. Otýpková & Chytrý (2006) analysed grassland and forest data sets and found the ordinations were similar between abundance and presence/absence analyses for some data sets, but rather different for others.

The overall situation is not clear, although Otýpková & Chytrý suggested that abundance information would be useful in homogeneous data sets.

The informativeness of presence/absence and abundance data

Demonstration that the presence/absence- and abundance-based analyses are often similar, but sometimes different, does not tell us which is closer to the intrinsic community structure, that is, which is closer to the assembly rules that govern community assembly, which better represents response to the environment, or which is more robust to sampling variation. There has been scarcely any direct investigation of this question. Lambert & Dale (1964) report the analysis of 15 'rather similar' heathland sites using the method of Williams & Dale (1962), partitioning the information content (i.e. the correlations between species) into separate correlation matrices for the qualitative and purely quantitative (i.e. abundance-when-present) data. They found that not only was there little extra information in the abundance-when-present element, but that the presence/absence data contained substantially more information than the presence/absence and abundance-when-present data combined. The only ecological interpretation of this has been that with abundance data, if the species is present its abundance will be an indication of how suitable the site is for it, but if the species is absent there is no information on how unsuitable it is. As Dale et al. (2001) commented: 'we cannot measure how much a species is absent; some noughts are "noughtier" … than others.' They described abundance-when-absent as 'a missing value.' Dale et al. (2001) partitioned information by determining the minimal message length needed to transmit a non-hierarchical clustering, optimizing message length with a quadratic likelihood function. They

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

concluded for one data set that the amount of information in the abundance-when-present data was slightly less than that in the presence/absence data. They commented: 'Again, we have a reminder that much of the pattern in vegetation data is related to patterns of absence' (Dale et al. 2001).

Dale and co-workers established the theory (Williams & Dale 1962; Lambert & Dale 1964; Dale et al. 2001), but meanwhile a vast number of vegetation surveys have been carried out, with time spent assessing the abundance of each species, perhaps because it is so counter-intuitive that presence/absence data could be as informative as, or more informative than, abundance data. The true test of this must be not through mathematics, but with real communities in the field. The problem in comparing analyses is that when a presence/absence and an abundance analysis give different results, which best reflects the structure of the community, and which is the most useful practically? A definitive answer is impossible, but I suggest two partial answers: habitat correlation and stability under subsampling. Both could be asked via a classification of the quadrats (cf. Feoli & Orloci 1979), but I here use DCA ordination because of the simplicity and intuition of comparing one ordination axis with another using a correlation coefficient. The answer may depend on the quality of the abundance measure, and this can be investigated not only by correlation across data sets but also by transforming the data.

These issues are investigated here on ten real-world data sets.

Habitat correlation

If a vegetation analysis correlates with known environmental factors, the result clearly represents some ecological reality. If the analysis has no environmental correlation, either the analysis is related to present or past environmental factors not measured, or it is less informative about the structure of the community. This criterion assumes that the distribution of species is largely controlled by the environment, one of the principles of the 'similarity theory' of Feoli & Orloci (2011): 'The first and most important [paradigm class] is that similar plant communities correspond to similar physical–chemical environments and should be convergent.' The question can be meaningfully asked only if the data set contains fairly full information on the environmental variation.

One practical study has addressed this issue, that of Smartt et al. (1976) with heathland data. They found that classification of the quadrat/species information using presence/absence data matched the environmental variation better than using any abundance measure. However, we should not conclude too much from one study, especially one mainly restricted to soil data.

Stability under subsampling

It is hardly ever possible to completely record the vegetation of an area. Instead, we rely on quadrats, hoping they will give a reliable estimate of the whole. An analogy for this, in fact a perfect replica, is a random subset of the quadrats that we have sampled. If the analysis is stable then analyses we make on a subset should be similar to those on the whole data set, in the same way that we always hope that the result of analysing our quadrats will be stable against the unrealized possibility of sampling another set, or the whole area.

Again, the concept is that such stability indicates that the data set may be representing some reality about the structure of the community. If the analysis is not stable under subsampling it seems likely that analysis of the whole presence/absence or abundance data set is not a stable representation of that reality.

Categorical scales of abundance

The data sets used here differ in the quality of the abundance information, as do data sets across the world. The +, 1–5 Braun-Blanquet 'cover–abundance' scale is widely used (Kershaw 1985). It is unsuited for standard multivariate analyses because it is non-metric (Podani 2005). However, another problem with such scales is that reduction to categories must lose information, as do categories of subjective cover (e.g. Peet et al. 1998). To judge the extent of the loss, and whether a categorical scale is better than presence/absence, analyses were repeated for three sites with the data quality reduced to a categorical scale. Field assignment to a categorical scale is always subjective and there must therefore be errors, so jittering was used on the same three data sets to judge the effect of such errors on the analyses.

Methods

Data sets and analysis

Ten data sets from southern New Zealand were selected to comprise a range of spatial extents (from 235 km down to 10 m), of habitat types (including carr, heathland, tussock grassland and short grassland), of environments (including sub-antarctic, mesic, bog and semi-arid) and of quadrat sizes (1000 m^2 down to 0.01 m^2) (Table 1, Appendix S1a). They were all data sets for which I could personally vouch for the sampling quality, but included measures of abundance with a range of accuracy in terms of representing the actual abundance of each species (Wilson 2011): height (very poor), subjective–ocular (possibly accurate but intrinsically open to question), two-dimensional frequency (objective, but imperfectly correlated with biomass), height–frequency (objective, and an approximation

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

Table 1. The data sets used. For further details see Appendix S1a. Gradient length is in SD units (Hill & Gauch 1980) from a DCA ordination based on the abundance data. Florule/richness = number of species in the data set/mean quadrat richness.

Data set	Extent	Habitat/community	Quadrats	Number of quadrats	Measure of abundance	Gradient length			Reference
						Axis 1	Axis 2	Florule/ richness	
Tussock landscape	235 km	Sub-alpine/alpine grasslands	1000 m^2, sampled with a 100 m line of 100 cm^3 quadrats, up to the height of the vegetation	264	Height frequency	4.91	4.92	6.65	Wilson & Meurk (2011)
Dryland landscape	40 km	Agricultural, indigenous grassland and shrubland	1 × 3 m	95	Ocular abundance on a 3-point scale	6.49	3.53	9.62	Wilson et al. (1989)
Sub-antarctic landscape	16 km	Grassland/herbfield/ shrubland/low forest	Variable area	140	Biomass index, calculated from ocular estimates of height, cover and frequency	8.41	4.57	7.59	Meurk et al. (1994)
Bog	3 km	Ombrotrophic bog and surrounds	1-m diameter	376	Ocular estimate of cover, converted to ranks	7.21	5.78	13.52	McQueen & Wilson (2000)
Dryland Hill	2 km	Semi-arid grassland	Four 0.5 × 0.5 m	114	Local shoot frequency (100 sub-squares)	5.91	3.62	9.83	Walker et al. (1995)
Carr	140 m	Tree-covered fen	5 × 5 m	64	Height	5.31	2.89	8.63	Sykes et al. (1991)
Heathland	50 m	Open grassland with small trees	Random 0.5 × 0.5 m	50	Local shoot frequency (25 sub-squares)	3.95	2.04	3.94	Matsui et al. (2002)
Lawn	10 m	Mesic lawn	Random 0.5 × 0.5 m	40	281 points per quadrat, all hits recorded	1.67	1.26	1.81	Roxburgh & Wilson (2000)
Dryland Microscale A	25 m	Dry-climate herbfield, selected for uniformity	Contiguous 0.1 × 0.1 m	256	Above-ground biomass	3.06	2.35	3.54	Wilson et al. (2000)
Dryland Microscale B	25 m	Semi-arid herbfield, selected for uniformity	Contiguous 0.1 × 0.1 m	256	Above-ground biomass	2.01	3.33	3.10	Wilson et al. (2000)

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

to biomass), repeated cover (objective, and an excellent measure of the contribution of a species to the community) and biomass (the ideal measure).

The vegetation analysis used detrended correspondence analysis (DCA) ordination (Hill & Gauch 1980; Appendix S1b), chosen because: (1) several studies have found that for the purpose of extracting a community gradient it is more efficient than methods such as NMDS (Ejrnaes 2000; Ruokolainen & Salo 2006; Hirst & Jackson 2007), (2) it typically yields two or more ecologically interpretable axes (Økland 1990; Table 2), its efficiency allowed many replicate random subsets to be run, and (3) it provides an estimate of gradient length (as SD units).

Habitat correlation

Four of the studies recorded good environmental information for each quadrat; all four sampled on a landscape extent. Because the studies were originally separate, and because the probable controlling factors differed between areas, different sets of environmental factors were measured (Table 3).

Quadrat scores on DCA axes were related to the environmental factors by multiple linear regression, separately for each axis 1 and separately using ordinations of the abundance data and of the data reduced to species presence/absence. Quadratic regressions were not used for the main analyses because of the danger of over-fitting with many environmental variates included already, but these are reported for the record.

Stability under subsampling

All ten data sets were used for this test. For each data set, 10 000 random subsets of quadrats were taken at several fractions of the complete number of quadrats: from 1.0 (the full data set), fractions down in 0.1 increments to 0.3, then down in 0.05 increments to a minimum of six quadrats. For each fraction, for each of the 10 000 random subsets, a Pearson product-moment correlation coefficient was calculated between the first DCA ordination axis of that subset and that using the whole data set, using only quadrats present in the subset. The absolute correlation coefficients were then meaned over the randomizations of that fraction (Appendix S2). This was done separately using the collected abundance data, and also with data reduced to presence/absence. Higher stability for abundance analyses or for presence/absence analyses would result in a higher correlation coefficient.

Categorical scales of abundance

For the sites with the highest-quality abundance data (Lawn, Dryland Microscale A and B), additional analyses were performed after conversion of the abundances to a 1–5 log scale of categories (Appendix S1c), somewhat comparable to the Braun-Blanquet scale often used in vegetation surveys (Greig-Smith 1983). A further analysis was done with jittering to simulate recording errors in the field, by at random increasing or decreasing the category of a random 25% of the quadrat/species records (but keeping within the 1–5 range): Appendix S1d.

Results

Habitat correlation

Presence/absence analyses matched the environmental factors better than abundance analyses in seven cases out of eight (Table 2). In some cases, the difference was substantial, all of them where the presence/absence ordination gave the closer fit to the environment. Overall, the fit to presence/absence was closer by +5.9%. A similar pattern was seen with Axes 3 and 4 (Appendix S1e), with the presence/absence fit closer on average by +6.8%. With quadratic regression (Appendix S1e) the advantage of presence/absence was not so clear-cut, but the three largest differences were all in that direction, as was the overall mean (+3.2%). Significance is not the issue here, but for the record for all regressions $P < 0.001$.

Stability under subsampling

For the first six data sets, those over larger extent, for almost every fraction (i.e. size of subset), the presence/absence analyses were at least as stable as the abundance ones, i.e. with a higher correlation coefficient (Fig. 1 a–f),

Table 2. Percentage of variation in quadrat DCA axes 1 and 2 scores accounted for by the environmental factors. For each pair, presence/absence vs abundance, the higher value is in bold.

Data set	Axis 1			Axis 2		
	Presence/absence	Abundance	Advantage to P/A	Presence/Absence	Abundance	Advantage to P/A
Tussock Landscape	**77.2**	62.4	+14.8	**72.1**	62.1	+10.0
Dryland Landscape	72.0	**72.3**	−0.3	**74.7**	73.4	+1.3
Sub-antarctic Landscape	**88.3**	87.7	+0.6	**86.1**	75.0	+11.1
Dryland Hill	**41.4**	40.7	+0.7	**37.6**	28.9	+8.7

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

Table 3. The environmental factors measured in the four landscape-scale studies used for correlation with habitat.

Tussock landscape	Dryland landscape	Sub-Antarctic landscape	Dryland hill
Climate	Position	Climate	Topography
Altitude (quadratic)	Latitude (paralleling the rainfall gradient)	Altitude	Altitude (quadratic)
Insolation	Longitude	Slope	North aspect
East aspect	Elevation	Insolation	East aspect
Slope	Topography	Distance to open coast	Slope
Rainfall	Exposure, 50-m scale	Distance to any coast	Insolation
Soil physical	Exposure, within the quadrat	Shelter	Soil physical
Soil depth	Slope	Wind speed	Topsoil depth
Bare rock and soil	North aspect	Soil physical	Total soil depth
Soil chemistry	East aspect	Temperature	Water
pH	Soil physical	Water	Organic content
Available Ca	Rock cover	Soil depth	Clay%
Available K	Stone (>2 mm)	Soil chemical	Gravel + sand%
Available PO_4	Coarse sand (0.2–2.0 mm)	pH (field and laboratory)	Soil chemical
Grazing (faecal pellets)	Fine sand (0.02–0.20 mm)	Available Ca	pH
	Silt (0.002–0.020 mm)	Available Mg	Available K
	Clay (<0.002 mm)	Available Na	Available Na
	Water	Available K	K/Na
	Soil chemical	Salinity (as conductivity)	Grazing (rabbit faecal pellets)
	pH	Grazing (faecal pellets and plant height)	
	Total N		
	Available PO_4		
	Available K		
	Available Mg		
	Available Ca		
	SO_4		
	C		
	Bioassay (growth of *Lolium perenne* in greenhouse)		
	Grazing (faecal pellets)		

considerably more stable for the Tussock Landscape and Carr. There were anomalous patterns at very low fractions, probably because with small subsets some correlations were indeterminate, and so were excluded from the means. However, for the Heathland, Lawn and two Dryland Microscale communities the result was opposite: greater stability under subsampling with abundance data.

Categorical scales of abundance

After conversion of the abundances to a 1–5 scale, stability with abundance data decreased dramatically for Dryland Microscale A, losing any advantage over presence/absence. Stability decreased also for the Lawn and Dryland Microscale B, although not to the level of presence/absence. Jittering decreased stability slightly for Dryland Microscale A, but had no detrimental effect for the other two sites.

Discussion

The results are mixed, with closer correlation with the environment for presence/absence in most cases, but not in all, and with greater stability under subsampling for presence/absence in some cases, but not in others. It is perhaps surprising that from the two criteria used here, abundance data sometimes gave an inferior result. Abundance data must incorporate presence/absence data, but it seems they can be misleading, probably because the zero values are ambiguous in that they give no indication of how far from the species ecological range the site is (Dale et al. 2001). However, abundance information is sometimes valuable; we need to examine under what circumstances.

Are the criteria used here appropriate?

The environmental correlations are open to the objection that the controlling factor might be one that was not measured, or could be a historical disturbance no longer observable. However, I chose only data sets where the environmental information was reasonably complete for the region being sampled. Occam's razor suggests that we should prefer the analysis that matches the environmental factors that we have, rather than one that might match hypothetical, unmeasured factor(s). Ecologists differ in their preference for ordination methods. In particular, it has been claimed that the second axis of a DCA ordination is distorted,

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

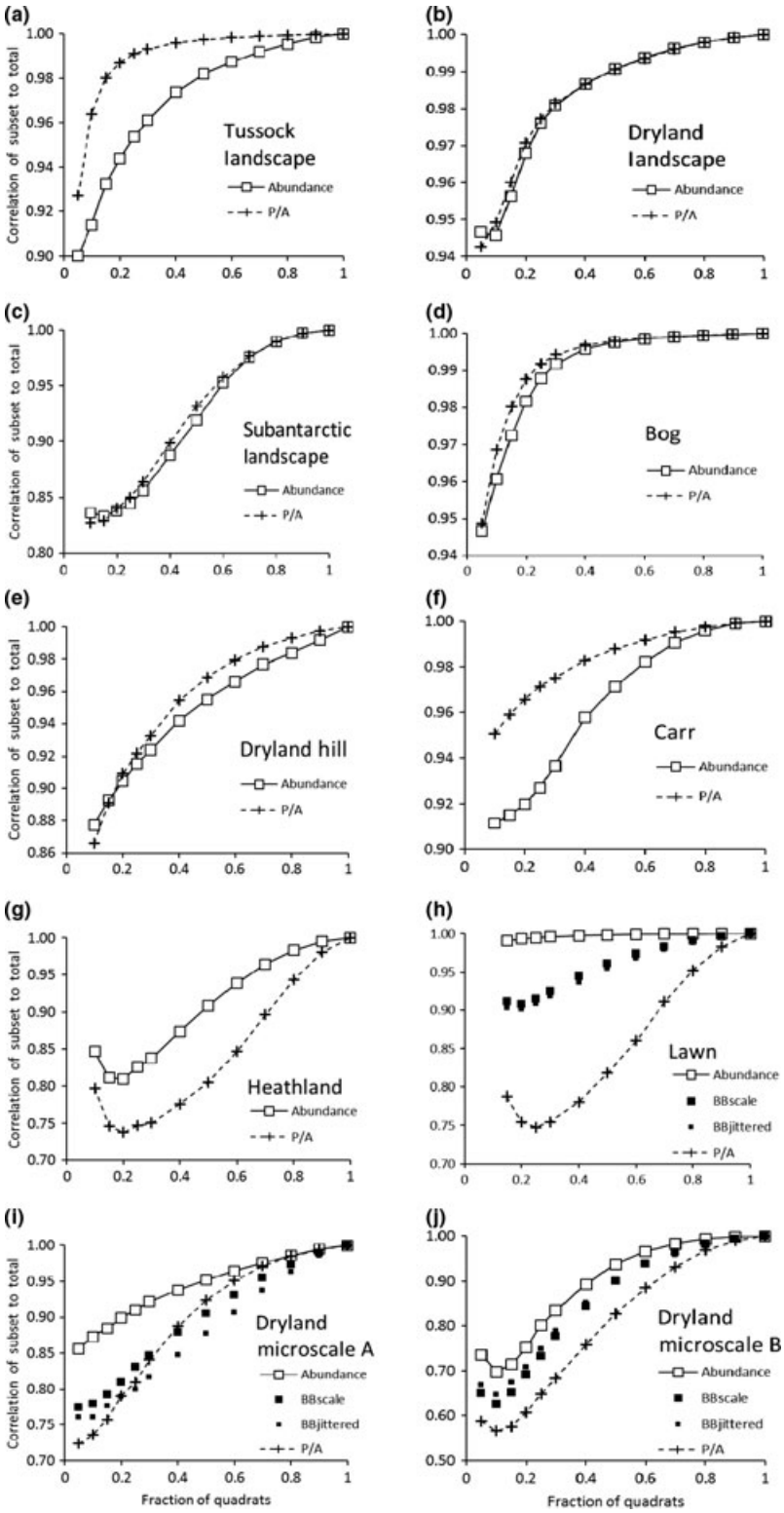


Fig. 1. For ten sites, correlation between DCA Axis 1 ordination scores from the full data sets and from those randomly subsampled to various fractions, the correlation using only the quadrats present in the subsample. Analyses were performed using the original abundance data, and with the data reduced to presence/absence. For three sites, results are shown with the abundance data reduced to categories ('BBscale'), and with the latter data jittered ('BBjittered').

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

and difficult to interpret ecologically (e.g. Minchin 1987). The present results (Table 2) refute this: although Axis 2 scores tend to be less related to environmental factors, here the percentage explained for Axis 2 (mean 63.7%) is only slightly less than for the major axis (mean 67.7%).

Subsampling, perhaps at first sight an arbitrary criterion, is in fact the most apposite, since in all vegetation sampling, the quadrats represent a small fraction of the total area, and the question is how well our sample reflects the reality of the whole. Subsampling the existing data set is an exact analogy of this.

When is abundance more informative?

The Heathland, Lawn and the two Microscale Dryland data sets (Fig. 1 g–j) differed from the first six (Fig. 1 a–f) in having greater stability with abundance data. There are four differences between these two groups: extent, heterogeneity, disturbance and data quality.

Extent

The four data sets showing greater stability with abundance were over smaller extents: at the extremes 10 m for the Lawn compared to 235 km for the Tussock Landscape study (Table 1, Fig. 1). However, whilst the Tussock Landscape, with the greatest extent, showed a clear advantage of presence/absence, so did the heterogeneous Carr data set with an extent of only 140 m. Effects of extent can be seen also in relations with the environment, with lower percentages explained for Dryland Hill over a much smaller extent. Extent may be a proxy for heterogeneity to some degree, but there were considerably closer environmental relations for the presence/absence analysis in the large-extent Tussock Landscape, even though the vegetation gradient was not especially long (Tables 1, 2).

Heterogeneity

The four data sets showing abundance advantage (Fig. 1 g–j) were all selected as uniform in habitat and vegetation, for studying community structure (texture convergence, relative abundance distributions, guild proportionality, etc.). This is reflected in their shorter gradient lengths (Table 1). In contrast, the first six studies (Fig. 1 a–f) were intended to distinguish different vegetation types out of known heterogeneity, for example the Bog study, which comprised bog and surrounding grassland/shrubland with completely different suites of species, and the Carr, which included saltmarsh, grassland, carr and terrestrial forest.

It has been suggested before that using abundance information will make less difference to the results of multivariate analyses when β-diversity is high (Økland 1990; Otýpková & Chytrý 2006). However, the latter work does not indicate which is preferable ecologically, and the present results show that: (a) the more relevant result is not necessarily the abundance-based analysis, and (b) the relation to heterogeneity is not clear-cut, e.g. the two data sets showing clear advantage for presence/absence (Tussock Landscape and Carr) were not among the four with the greatest DCA Axis 1 gradient lengths.

Disturbance

The two Dryland Microscale sites were identical in sampling method (256 contiguous quadrats with above-ground biomass), of similar physiognomy (short grass/forb/cushion) and 40-km apart in the rain shadow of the Southern Alps. Yet they differed. Site A, with subsampling abundance stability down to 0.86 and presence/absence stability down to 0.72 was under about 650 $mm \cdot yr^{-1}$ rainfall, with a significant native species component. The community was probably at equilibrium when sampled. Dryland Microscale B was under 400 $mm \cdot yr^{-1}$ rainfall, and stability under subsampling was lower, reaching a minimum with abundance data of 0.69 and with presence/absence 0.57. The site was disturbed and the species mainly exotic, so we might speculate that the abundances of species had not reached equilibrium and were less predictable.

Data quality

The fourth difference is that the four data sets showing an advantage of abundance all have high-quality data (the issue of quality of presence/absence data hardly arises here, since temporary false absences at small grain are hardly a problem for sedentary plants, and in the data sets used here there was practically no problem of false absences of summer-green species). For the two Dryland Microscale sites, all above-ground plant material was taken, sorted to species, and biomass determined, i.e. the ideal measure of abundance (Wilson 2011). For the Lawn study, repeated cover was used (Wilson 2011) with a very fine needle point (<0.05 mm). The Heathland data were local shoot frequency in 25 sub-quadrats; this is an objective measure, although with the sub-quadrat size used there it gives an inferior estimate of the plant material present to biomass or repeated cover with point quadrats. The Carr was not so different in extent from the Heathland, but with poor-quality abundance data, and presence/absence gave greater stability.

However, the relation is not clear-cut: the data quality for the Tussock Landscape was high: local frequency, across 100 positions, with a small quadrat, and with quadrat volumes extending up to the top of the canopy giving 3-D frequency, i.e. the closest approach of any local

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

frequency measure to biomass. Yet it showed the clearest advantage for presence/absence, in both habitat correlation and stability under subsampling, presumably due to its greater extent and heterogeneity in environment and in past disturbance (Wilson & Meurk 2011).

Categorical scales of abundance

There are various category scales used for abundance, several standard ones and many ad hoc; however, a 1–5 scale is typical. The conversion to a 1–5 scale for the Lawn and for Dryland Microscale sites makes it clear that at least some of the value of abundance information (Fig. 1h,j), sometimes all (Fig. 1i), is due to the precision of the abundance measurement and is lost in conversion to categories. The use of a categorical scale is often justified on the basis that it largely overcomes observer effects in subjective estimates, but reduction to categories never increases the accuracy of data, and can only result in information loss. In fact, jittering showed that random errors in recording have only a minor effect, a slight effect for Dryland Microscale A but no effect at all for the other two data sets. Of course, errors resulting from some bias, such as the demonstrated increase in subjective ratings for herbs when they are flowering (Greig-Smith 1983), will have considerably more effect than these random ones.

When are abundance data more informative? Conclusions

There is some correlation between attributes among the data sets used here, in that the three with smallest extent and least heterogeneity have the highest-quality data, although the two with greatest extent have higher-quality data than the intermediate ones. However, the analyses where abundance was reduced to categories demonstrate that much or all of the value of abundance information is lost in this form. From the results available, the suggestion is that abundance data are of value only when sampling in over a small extent and in homogeneous communities and when high-quality information is taken, such as biomass or repeated cover (i.e. point quadrats with every hit recorded). Probably, homogeneity and data quality are the more important.

Practical conclusions

Environmental correlation is sometimes the aim of sampling, and even in other cases vegetation surveyors would normally prefer a result that correlated with known environmental factors than having to speculate that it is due to unknown, unmeasured factors. For this, presence/absence is generally as good, and usually better (Table 2). Others have found the presence of species to be a better indicator of the environment than any abundance-based measure, for example the mollusc *Dreissena bugensis* in the study of Podani & Csány (2010). Marignani et al. (2008) found that when the data were reduced to life form or growth form, presence indication was effective for discriminating between prior land-cover classes, whereas abundance-based indicators were not.

Presence/absence will also be more useful if indicator species are required for subsequent field identification of the associations formed (e.g. the application discussed in Dufrêne & Legendre 1997). Surveyors will also want a result that is not too dependent on how many quadrats are taken and just where they are placed, and for this presence/absence sampling should be used, being as stable as abundance or more so, and of course giving faster sampling (Fig. 1 a–f). Williams et al. (1973), considering a tropical rain forest data set, and with their only criterion being whether the result seemed 'ecologically interpretable' in terms of structural complexity and occurrence on soil type, concluded that for sampling large areas 'the results … are not impaired' by including abundance information. I have demonstrated impairment in some data sets. Abundance can be misleading even when measured without error, being affected by season of sampling, year-to-year variation, grazing, etc.

The overall practical conclusion is clear. For vegetation surveys including the range of variation typically found across an extent of more than 100 m (and most are over many kilometres or even thousands of kilometres), there is no advantage in collecting abundance data in any of the forms that are commonly used: cover by-eye, subjective categories, or even local frequency. Presence/absence is quicker to collect, sufficient and may better represent the communities of the area.

The subsampling approach also tells us retrospectively whether there were enough quadrats; a high correlation with the full data set obtained with a fraction of the quadrats tells us that fraction would have been adequate for defining community types (cf. Pillar 1998). For the Tussock Landscape, Dryland Landscape and Bog studies, the gradient in species composition would have been clear with only half the number of quadrats used ($r \geq 0.98$), at least with presence/absence, although the number used were needed in order to make maps of the distribution of community types. On the other hand, the Dryland Hill would probably have been better characterized by more than the 140 quadrats taken.

Theoretical considerations

The result also has implications for the structure of communities, especially since under most analyses abundance data imply a much greater proportional contribution of the minor

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

species (Smartt et al. 1976). The superiority of presence/absence for the larger-extent data sets here supports the theory behind the widespread approach of the Zürich-Montpellier school, that the presence of 'character' species (constant/faithful/differential/diagnostic), often quite minor ones, is the best indicator of community type (Braun-Blanquet 1932). Indeed, for the Tussock Grassland data set used here, Wilson & Meurk (2011) found that the environment related more closely to distribution of subordinate species than to the single dominant, and little less than to the group of dominants. On the other hand, Webb et al. (1967) found that big tree species best indicated tropical rain forest community composition in Queensland, Australia, and the classification of North American vegetation of Weaver & Clements (1929) was based on the 'dominants or controlling species.'

In the considerable discussion on assembly rules, that limit the occurrence and abundance of species in a community (Wilson 1999, 2007; Grime 2006; Mouchet et al. 2010), there have been suggestions that assembly rules exercise tighter controls on the occurrence of species in a community than on their abundance. There is some evidence for this (Lawrence Lodge et al. 2007), and it is supported here for greater extents than those covered in Lawrence Lodge et al.

Acknowledgements

I thank the many colleagues who have helped me to collect the data used here: Ralph Allen, Nathan Dougherty, Martin Foggo, Warren McGeorge King, Barry Laurence, Bill Lee, Kelvin Lloyd, Abi Loughnan, Alan Mark, Tetsuya Matsui, Amelia McQueen, Colin Meurk, Jamie Newman, John Steel, Jo Swaney, Susan Walker and Peter Williams. For comments on a draft, I thank Milan Chytrý, Mike Dale, Enrico Feoli, Rune Halvorsen, Jill Hetherington, János Podani, Cailin Roe and Ed Wilson.

References

Avena, G., Blasi, C., Feoli, E. & Scoppola, A. 1981. Measurement of the predictive value of species lists for species cover in phytosociological samples. *Vegetatio* 45: 77–84.

Braun-Blanquet, J. 1932. *Plant sociology: the study of plant communities*. McGraw-Hill, New York, NY, US.

Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C. & Lane, A.M.J. 1996. ITE Merlewood land classification of Great Britain. *Journal of Biogeography* 23: 625–634.

Dale, M.B. 1995. Evaluating classification strategies. *Journal of Vegetation Science* 6: 437–440.

Dale, M.B., Salmina, L. & Mucina, L. 2001. Minimum message length clustering: an explication and some applications to vegetation data. *Community Ecology* 2: 231–247.

Dufrêne, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* 67: 345–366.

Ejrnaes, R. 2000. Can we trust gradients extracted by Detrended Correspondence Analysis? *Journal of Vegetation Science* 11: 565–572.

Feoli, E. & Orloci, L. 1979. Analysis of concentration and detection of underlying factors in structured tables. *Vegetatio* 40: 49–54.

Feoli, E. & Orloci, L. 2011. Can similarity theory contribute to the development of a general theory of the plant community? *Community Ecology* 12: 135–141.

Feoli, E., Ferro, G. & Ganis, P. 2006. Validation of phytosociological classifications based on a fuzzy set approach. *Community Ecology* 7: 99–108.

Greig-Smith, P. 1983. *Quantitative plant ecology*, 3rd ed. Blackwell, Oxford, UK.

Grime, J.P. 2006. Trait convergence and trait divergence in herbaceous plant communities: mechanisms and consequences. *Journal of Vegetation Science* 17: 255–260.

Hill, M.O. & Gauch, H.G. 1980. Detrended correspondence analysis, an improved ordination technique. *Vegetatio* 42: 47–58.

Hirst, C.N. & Jackson, D.A. 2007. Reconstructing community relationships: the impact of sampling error, ordination approach, and gradient length. *Diversity and Distributions* 13: 361–371.

Jennings, M.D., Faber-Langendoen, D., Loucks, O.L., Peet, R.K. & Roberts, D. 2009. Standards for associations and alliances of the US National Vegetation Classification. *Ecological Monographs* 79: 173–199.

Kershaw, K.A. 1985. *Quantitative and dynamic plant ecology*. Edward Arnold, London, UK.

Lambert, J.M. & Dale, M.B. 1964. The use of statistics in phytosociology. *Advances in Ecological Research* 2: 59–99.

Lawrence Lodge, R.H.E., Anderson, B.J., de Groot, A., Bill, A., McQueen, A.A.M., Steel, J.B., Mistral, M., Mason, N.W.H. & Wilson, J.B. 2007. Spatial autocorrelation in plant communities: vegetation texture versus species composition. *Ecography* 30: 801–811.

Marignani, M., Delvico, E. & Maccherini, S. 2008. Performance of indicators and the effect of grain size in the discrimination of plant communities for restoration purposes. *Community Ecology* 9: 201–206.

Matsui, T., Dougherty, N.J., Loughnan, A.E., Swaney, J.K., Laurence, B.L., Lloyd, K.M. & Wilson, J.B. 2002. Local texture convergence within three communities in Fiordland, New Zealand. *New Zealand Journal of Ecology* 26: 15–22.

McDonald, D.J. 1997. VEGMAP: a collaborative project for a new vegetation map of southern Africa. *South African Journal of Science* 93: 424–426.

McQueen, A.A.M. & Wilson, J.B. 2000. Vegetation and environment of a New Zealand raised bog. *Journal of Vegetation Science* 11: 547–554.

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science

Meurk, C.D., Foggo, M.N. & Wilson, J.B. 1994. The vegetation of subantarctic Campbell Island. *New Zealand Journal of Ecology* 18: 123–168.

Minchin, P.R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio* 69: 89–107.

Mouchet, M.A., Villeger, S., Mason, N.W.H. & Mouillot, D. 2010. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology* 24: 867–876.

Økland, R.H. 1990. Vegetation ecology: theory, methods and applications with reference to Fennoscandia. *Sommerfeltia supplement* 1: 1–233.

Orloci, L. 1966. Geometric models in ecology: I The theory and application of some ordination methods. *Journal of Ecology* 54: 193–215.

Otýpková, Z. & Chytrý, M. 2006. Effects of plot size on the ordination of vegetation samples. *Journal of Vegetation Science* 17: 465–472.

Peet, R.K., Wentworth, T.R. & White, P.S. 1998. A flexible, multipurpose method for recording vegetation composition and structure. *Castanea* 63: 262–274.

Pillar, V.D. 1998. Sampling sufficiency in ecological surveys. *Abstracta Botanica* 22: 37–48.

Podani, J. 2005. Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. *Journal of Vegetation Science* 16: 497–510.

Podani, J. & Csány, B. 2010. Detecting indicator species: some extensions of the IndVal measure. *Ecological Indicators* 10: 1119–1124.

Roxburgh, S.H. & Wilson, J.B. 2000. Stability and coexistence in a lawn community: experimental assessment of the stability of the actual community. *Oikos* 88: 409–423.

Ruokolainen, L. & Salo, K. 2006. Differences in performance of four ordination methods on a complex vegetation dataset. *Annales Botanici Fennici* 43: 269–275.

Schaminée, J.H.J., Hennekens, S.M. & Ozinga, W.A. 2007. Use of the ecological information system SynBioSys for the analysis of large data sets. *Journal of Vegetation Science* 18: 463–470.

Smartt, P.F.M., Meacock, S.E. & Lambert, J.M. 1976. Investigations into the proper ties of quantitative vegetational data. II. Further data type comparisons. *Journal of Ecology* 64: 41–78.

Sykes, M.T., Wilson, J.B., Lee, W.G. & Allen, R.B. 1991. A remnant New Zealand carr. *Journal of Vegetation Science*, 2: 209–216.

Vermeersch, S., De Genst, W., Vermoesen, F. & Triest, L. 2003. The influence of transformations of an ordinal scale of a floristic gradient, applied on a TWINSPAN classification. *Flora* 198: 389–403.

Walker, S., Mark, A.F. & Wilson, J.B. 1995. The vegetation of Flat Top Hill: an area of semi-arid grassland/shrubland in Central Otago, New Zealand. *New Zealand Journal of Ecology*, 19: 175–194.

Weaver, J.E. & Clements, F.E. 1929. *Plant ecology*, 1st ed. McGraw-Hill, New York, NY, US.

Webb, L.J., Tracey, J.G., Williams, W.T. & Lance, G.N. 1967. Studies in the numerical analysis of complex rain-forest communities. II. The problem of species-sampling. *Journal of Ecology* 55: 525–538.

Williams, W.T. & Dale, M.B. 1962. Partition correlation matrices for heterogeneous quantitative data. *Nature* 196: 602.

Williams, W.T., Lance, G.N., Webb, L.J. & Tracey, J.G. 1973. Studies in the numerical analysis of complex rain-forest communities. VI. Models for the classification of quantitative data. *Journal of Ecology* 61: 47–70.

Wilson, J.B. 1999. Assembly rules in plant communities. In: Weiher, E. & Keddy, P.A. (eds.) *Ecological assembly rules: perspectives, advances, retreats*. pp. 130–164. Cambridge University Press, Cambridge, UK.

Wilson, J.B. 2007. Trait-divergence assembly rules have been demonstrated: limiting similarity lives! A reply to Grime. *Journal of Vegetation Science* 18: 451–452.

Wilson, J.B. 2011. Cover plus: ways of measuring plant canopies and the terms used for them. *Journal of Vegetation Science* 22: 197–206.

Wilson, J.B. & Meurk, C.D. 2011. The control of community composition by distance, environment and history: a regional-scale study of the mountain grasslands of southern New Zealand. *Journal of Biogeography* 38: 2384–2396.

Wilson, J.B. & Stubbs, W.J. 2012. Evidence for assembly rules: limiting similarity within a saltmarsh. *Journal of Ecology* 100: 210–221.

Wilson, J.B., Williams, P.A. & Lee, W.G. 1989. Vegetation composition and segregation in relation to the environment at low altitudes in the Upper Clutha Basin, New Zealand. *New Zealand Journal of Ecology*, 12: 103–116.

Wilson, J.B., Steel, J.B., Newman, J.E. & King, W.Mc.G. 2000. Quantitative aspects of community structure, examined in a semi-arid grassland. *Journal of Ecology*, 88: 749–756.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1. (a) Further details of the data sets used and the analyses; (b) Program and data notes; (c) Conversion of abundances to a 1–5 scale; (d) Algorithm for jittering; (e) Axes 3 and 4 (linear regression); (f) Quadratic regression; (g) Eigenvalues.

Appendix S2. Correlation between subset and whole-data set ordination scores.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Doi: 10.1111/j.1654-1103.2012.01430.x © 2012 International Association for Vegetation Science