

Semi-automated assignment of vegetation survey plots within an *a priori* classification of vegetation types

Ian Oliver^{1,2,*}, Elizabeth A. Broese², Martin L. Dillon², Dominic Sivertsen³ and Megan J. McNellie⁴

¹School of Environmental and Rural Sciences, University of New England, Armidale, NSW, 2351, Australia; ²NSW Office of Environment and Heritage, PO Box U221, Armidale, NSW, 2351, Australia; ³NSW Office of Environment and Heritage, PO Box A290, Sydney South, NSW, 1232, Australia; and ⁴NSW Office of Environment and Heritage, PO Box 5336, Wagga Wagga, NSW, 2650, Australia

Summary

1. Assignment of large numbers of vegetation plots to *a priori* vegetation classifications is increasingly being required to support natural resource management, monitoring and conservation at regional scales. Several automated systems have been developed that use quantitative synoptic tables and algorithm-based plot-to-type assignment. However, where synoptic tables do not exist, and qualitative species lists characterise vegetation type classifications, existing systems may not apply. In these situations, vegetation experts may resort to manual assignment processes that can be slow, subjective and fraught with difficulties.

2. This study combines repeatable and objective quantitative analyses, with new software, to deliver a semi-automated plot-to-type assignment process appropriate for *a priori* classifications based on qualitative species lists. The flexible semi-automated assignment program (SAAP) calculates a quantitative goodness-of-fit score between plots and types, based on the species that characterise each *a priori* vegetation type, and the species that characterise groups of plots derived from quantitative analyses.

3. We applied the SAAP to a case-study of 630 native vascular plant species from 930 plots, and an *a priori* classification of 99 vegetation types. We varied vegetation data set transforms [cover per cent (0–100%), cover score (0–6) and presence–absence (1, 0)] and analysis settings and tested the degree to which the SAAP provided plot-to-type assignment concordant with manual expert assignment.

4. Results provided clear evidence supporting the choice of particular data set transformations and analysis settings to maximise concordance. The SAAP allocated up to 50% of plots to the same expert-assigned vegetation type, and more than 70% of plots to an expert-assigned vegetation type ranked in the top five by the SAAP.

5. When coupled with repeatable and objective quantitative analyses, the SAAP provides vegetation experts with a new semi-automated and quantitative decision support tool to assist with the assignment of vegetation plots within *a priori* vegetation classifications defined by characteristic species lists.

Key-words: semi-automated system, decision support, characteristic species, consistency, goodness-of-fit, numerical classification, relevés, vegetation classification

Introduction

Approaches used to group vegetation plots (relevés) to create floristically derived vegetation classifications are many and varied, and a source of much research and debate (Tichý *et al.* 2010; Chytrý, Schaminée & Schwabe 2011). This is not surprising when the objective of classification is to identify and delineate discrete communities, yet vegetation scientists mostly agree that, ‘vegetation types are abstract entities that delimit and name parts of the vegetation continuum to facilitate communication about them’ (De Cáceres & Wiser 2012; p. 387). Traditionally, identification and delineation of types (syntaxa, see Mucina 1997) along this continuum was based on expert knowl-

edge, and the resultant classifications are still widely accepted in many countries (Mucina 1997; Kočí, Chytrý & Tichý 2003; De Cáceres *et al.* 2009), including Australia (Benson *et al.* 2010). These *a priori* vegetation classifications are increasingly being used to support natural resource management, monitoring and conservation at large spatial scales (Jennings *et al.* 2009; Chytrý, Schaminée & Schwabe 2011; Schaminée *et al.* 2011).

Expert-based classifications can suffer from inconsistencies because the criteria for identification and delineation of vegetation types may vary among experts and are often not made explicit (Mucina 1997; Kočí, Chytrý & Tichý 2003). These problems can be overcome by applying formalised, consistent and repeatable quantitative methods to vegetation classification. However, there are a plethora of quantitative methods available (see McCune & Grace 2002; Tichý 2002; Kočí, Chytrý & Tichý 2003; van Tongeren, Gremmen & Hennekens

*Correspondence author. E-mail: ian.oliver@environment.nsw.gov.au

2008; De Cáceres, Font & Oliva 2010; Chytrý, Schaminée & Schwabe 2011), with each delivering potentially different classifications (De Cáceres, Font & Oliva 2010; Wiser *et al.* 2011). Even sophisticated combinations of methods have not received universal acceptance, especially in those European countries where phytosociological tradition is well established (Kočí, Chytrý & Tichý 2003; Schaminée *et al.* 2011). Achieving standardisation and consistency in vegetation classification remains an important goal for vegetation scientists (Mucina 1997; Anderson *et al.* 1998; Grossman *et al.* 1998; Rodwell 2006; Jennings *et al.* 2009).

De Cáceres & Wiser (2012) provide a conceptual framework for achieving consistency in vegetation classification centred on two ideas: (i) the need to explicitly distinguish between the conceptual activities involved in the definition of vegetation types, including the process of 'type characterisation' and (ii) the need to perform assignments of new vegetation observations to *a priori* vegetation types in accordance with how the types were originally defined, referred to as 'consistency in assignment'. Consistency in assigning new observations to an *a priori* classification, the focus of this study, is inseparable from characterisation, as it is the process of characterisation that defines the attributes of vegetation types to which the new observations will be compared. Our interest is in the identification and use of diagnostic or characteristic species for characterisation and consistency in assignment.

Where characterisation is based on quantitative methods, synoptic (or floristic) tables containing quantitative information for each diagnostic species for each vegetation type can be produced (Tichý 2002). Synoptic tables have the appeal of providing transparent, quantitative data, upon which assignment of new observations to an *a priori* classification can be based. However, synoptic tables can contain many different types of information. This is illustrated by the vegetation classification software JUICE (Tichý 2002) which can create synoptic tables using 'percentage constancy', 'categorical constancy', 'absolute constancy', 16 different fidelity measures based on either presence-absence or quantitative data, as well as a number of tables summarising measures of species cover or abundance among plots. Synoptic tables published for the UK National Vegetation Classification use 'categorical constancy' (referred to as frequency classes) and where the species is observed, the range in cover-abundance is recorded on a ten point Domin scale (see Rodwell 2006; fig. 12). In comparison, the US National Vegetation Classification uses 'percentage constancy', and where the species is observed, the range in percentage cover is recorded, as is the average percentage cover for all species from all plots within the vegetation type (Jennings *et al.* 2009; appendix D). To the authors' knowledge there is no European standard for the synoptic tables that underpin the various vegetation classifications developed within the European Union (but see Gégout & Coudun 2012).

Despite the lack of a synoptic table standard, several automated systems have been developed to consistently assign new observations to *a priori* classifications using synoptic table data. The software ASSOCIA (van Tongeren, Gremmen & Hennekens 2008) makes assignments using a composite index

of dissimilarity (combining qualitative and quantitative data) calculated between new plots and synoptic tables (of the form, percentage constancy, and average cover of species). ASSOCIA therefore aligns with the synoptic tables supporting the US National Vegetation Classification. Within the UK, the software TABLEFIT (Hill 1989, 1996) and modular analysis of vegetation information system (MAVIS, <http://www.ceh.ac.uk/products/software/index.html>) make assignments based on a modified Czekanowski index of similarity between new plots and the synoptic tables characterising the UK National Vegetation Classification (of the form, categorical constancy and minimum and maximum Domin abundance class). Individual researchers have also explored consistent assignment of plots to *a priori* classifications using a variety of supervised classification methods including artificial neural networks, fuzzy clustering and a variety of project specific matching indices (see De Cáceres *et al.* 2009; De Cáceres, Font & Oliva 2010; Gégout & Coudun 2012 and citations therein). However, as for the above automated systems based on synoptic tables, these methods first require a classifier developed from a training data set of plot observations, whose classification is already known and assumed to be valid (De Cáceres *et al.* 2009).

In parts of the world where phytosociological tradition is less well established, vegetation classifications and characterisations tend to be derived from expert knowledge, or derived only in part using quantitative methods (Kočí, Chytrý & Tichý 2003; De Cáceres *et al.* 2009; Benson *et al.* 2010). These classifications often lack formal synoptic tables or classifiers, with the floristic component of vegetation type characterisation limited to qualitative lists of diagnostic species. The usefulness of existing automated systems, based on synoptic tables, for assigning new vegetation observations to *a priori* classifications, may in these situations be limited. For example, in eastern Australia, at least three separate vegetation classifications exist: Ecological Vegetation Classes in Victoria (see <http://www.dse.vic.gov.au/conservation-and-environment/native-vegetation-groups-for-victoria>); Regional Ecosystems in Queensland (see Sattler & Williams 1999; http://www.ehp.qld.gov.au/ecosystems/biodiversity/re_introduction.html); and Plant Community Types in New South Wales (see Benson 2006; Benson *et al.* 2010; <http://www.environment.nsw.gov.au/research/Vegetationinformation/system.htm>). The Victorian classification system uses synoptic tables based on percentage constancy and a fidelity score, however, synoptic tables are developing only gradually for the Queensland classification and do not exist for the New South Wales classification. Completing formal synoptic tables for these very large and diverse classifications (1384 types in Queensland and an estimated 1500 types in New South Wales) will take many years. Characterisation of these types currently uses a general community description, biogeographic location and characteristic species lists.

In the absence of automated systems designed for these and other 'early-stage' classifications, vegetation scientists still need to resort to an expert-dependent manual sorting process, as was practiced in the 1960's (Mucina 1997), to assign vegetation observations to *a priori* classifications. These manual processes can be slow, subjective and fraught with difficulties (see van

Tongeren, Gremmen & Hennekens 2008; Willner 2011). This paper describes a flexible semi-automated process that combines objective and repeatable quantitative methods (numerical classification) with new software, the semi-automated assignment program (SAAP), to calculate quantitative goodness-of-fit scores between large numbers of vegetation plots and an *a priori* classification of vegetation types characterised by qualitative species data only. Using a case-study, we test the degree to which our semi-automated process and the SAAP, provides results concordant with manual expert assignment.

Materials and methods

Our study involved five steps: (1) manual assignment of vegetation plots to an *a priori* vegetation classification; (2) preparation of the *a priori* vegetation type characteristic species SAAP input file; (3) quantitative analysis of the vegetation plot data; (4) goodness-of-fit calculation between plots and types using the SAAP; and (5) comparison of SAAP results with the results of manual assignment (see Fig. 1).

STEP 1. MANUAL ASSIGNMENT OF VEGETATION PLOTS TO AN *A PRIORI* VEGETATION CLASSIFICATION

Vegetation plots were sourced from 11 separate vegetation surveys, undertaken within southern New South Wales, that consistently recorded cover and abundance (Appendix S1). The data set consisted of 630 native species from 930 plots. Manual assignment of plots to the *a priori* classification was undertaken by an experienced botanist familiar with the vegetation of the region (DS). The NSW Vegetation Classification and Assessment (VCA) database (Benson 2006; Benson *et al.* 2010) provided the study area's *a priori* classification of 99 VCA types for the region. These *a priori* vegetation types were characterised

according to lists of characteristic tree, shrub and groundcover species (field layer, see Jennings *et al.* 2009), derived from plot data analysis, or from field observations (see Benson 2006). Species recorded from each of the 930 plots were manually compared by the expert with the lists of characteristic species in the VCA database. Vegetation type was assigned to each plot based on dominance (cover and abundance) of characteristic species, together with a consideration of the plot's location, landscape position, and vegetation pattern shown on SPOT5 2.5 m or ADS40 0.5 m imagery. Results were checked qualitatively against existing mapping and classifications (Fig. 1, Step 1).

STEP 2. PREPARATION OF THE *A PRIORI* VEGETATION TYPE CHARACTERISTIC SPECIES SAAP INPUT FILE

The order that characteristic species were listed within each stratum (tree, shrub, groundcover) within the VCA database reflected the dominance of species derived from plot data analyses, or from field observations (J.S. Benson, personal communication). As is standard practice, this order is also reflected in the species used for the 'scientific name' attached to the VCA vegetation type (Benson 2006; Jennings *et al.* 2009). To account for this qualitative dominance data in SAAP analyses, ranks were applied to characteristic tree, shrub and groundcover species according to their order in the VCA database characteristic species fields. These ranks were incorporated into the *a priori* vegetation type characteristic species SAAP input file (Fig. 1, Step 2; Table S1) and used by the SAAP according to user-defined settings.

STEP 3. QUANTITATIVE ANALYSIS OF THE VEGETATION PLOT DATA

From the data set consisting of 630 native species from 930 plots, three data sets were prepared for quantitative analysis: (i) cover per cent (0–100%); (ii) cover-abundance score (0–6); and (iii) presence-absence

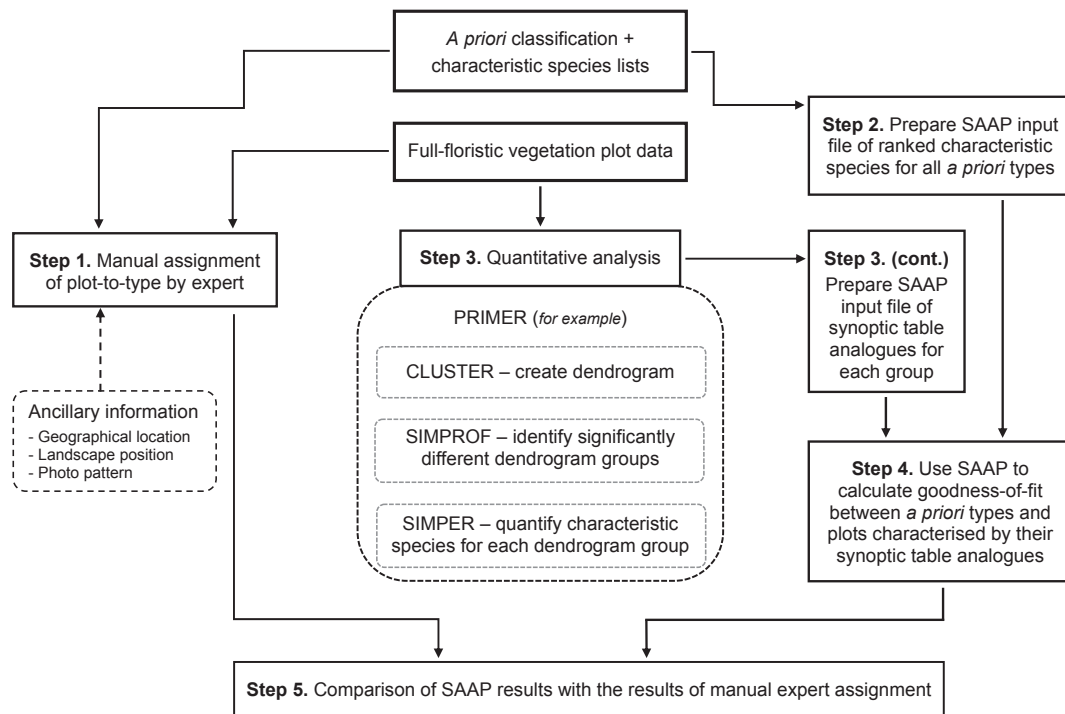


Fig. 1. Workflow showing the relationships between the steps undertaken for this study, and the positioning of the semi-automated assignment program, within our semi-automated process designed to assist experts assign vegetation plots to *a priori* vegetation types.

(1, 0); with the aim of exploring the influence of gradually reducing the effect of dominant species, within the analyses, on the concordance with the expert's results (see Dataset in Supporting Information for details on analysis data set preparation).

The three data sets were submitted to quantitative analysis using PRIMER V6 (Clarke & Gorley 2006). For each, a Bray–Curtis association matrix was used to generate a dendrogram, based on agglomerative hierarchical clustering and group average linkage. SIMPROF (similarity profiles, see Clarke, Somerfield & Gorley 2008) was used to provide an objective means of identifying significantly different dendrogram groups. To minimise Type I errors, alpha was reduced to 0.1% and statistical significance based on 9999 permutations. These analyses produced a robust, repeatable and objective numerical classification of plots for each of the three data sets and satisfied De Cáceres & Wiser's (2012) requirement to explicitly define the process of membership determination.

Group characterisation was achieved using the PRIMER routine SIMPER (similarity percentages, see Clarke & Gorley 2006). SIMPER quantified the degree to which species within plots, within each SIMPROF group, were characteristic of that group. SIMPER decomposes all similarities among plots, within a group, to quantify in descending order, the contribution each species makes to the average Bray–Curtis similarity within the group. Cut-off values of 90% and 99% were explored for these analyses, that is, no more species were listed once a cumulative contribution of either 90% or 99% was reached. Species with high contributions can be described as being characteristic, or diagnostic, of each of the objectively identified dendrogram groups. These robust, repeatable and objective approaches for membership determination and group characterisation produced quantitative data analogous to the synoptic tables that underpin the advanced vegetation classifications of the US and the UK. These 'synoptic table analogue' data were then prepared for input into the SAAP and used to quantify the goodness-of-fit between plots within dendrogram groups, and the *a priori* classification of vegetation types (Fig. 1, Step 3; Table S2).

STEP 4. THE SAAP'S GOODNESS-OF-FIT CALCULATIONS AND FLEXIBLE USER-DEFINED SETTINGS

The SAAP rapidly calculates a goodness-of-fit score between all *a priori* vegetation types, and all plots, via their dendrogram group synoptic table analogue data, generated by quantitative analysis (see Step 4). The goodness-of-fit score represents the sum of SIMPER contribution scores for each species recorded in each synoptic table analogue, that is also listed as a characteristic species in each available VCA type (see Dataset in Supporting Information for examples of goodness-of-fit score calculation).

The SAAP also provides flexibility to vary the number of VCA characteristic species considered, and to negatively or positively weight the SIMPER contribution scores associated with VCA characteristic species (see Dataset in Supporting Information for further discussion on species weighting). To explore the influence of this flexibility on assignment results, we undertook separate SAAP analyses for each of our three data set types (cover per cent, cover score and presence-absence), and for each of our two synoptic table analogue data sets (SIMPER cumulative contribution score cut-offs of 90% and 99%). With these six combinations, combined with the following user-defined settings, 108 separate SAAP analyses were performed:

- 1 Three variations on the maximum number of VCA type characteristic species per stratum considered by the SAAP (Top 3, Top 5, Top 10);
- 2 Three variations of down-weighting characteristic species according to their rank within the stratum (1.0, 0.75, 0.5), where smaller numbers resulted in less influence of lower ranked species;

- 3 Two variations of increased weighting applied to species from the dominant stratum of each vegetation type (1.0, 1.25), where larger numbers resulted in more influence of species from the dominant stratum (see Dataset in Supporting Information for further discussion of species weighting).

STEP 5. COMPARISON OF SAAP RESULTS WITH THE RESULTS OF MANUAL EXPERT ASSIGNMENT

Our manual expert-dependent plot-to-type assignment results were the benchmark against which our SAAP results were compared. To explore the goodness-of-fit of our 108 sets of SAAP results, with the manual expert-derived results, we tallied the number of plots for which the expert-assigned vegetation type was ranked first, second, third, up to tenth, by each SAAP analysis.

We explored the influence of: data set type; cumulative contribution cut-off level; and each of the user-defined settings within the SAAP; on the concordance with the expert's results, by averaging results for those analyses which kept one component constant. For example, the average number of times the expert-assigned vegetation type was ranked first by the SAAP using the cover per cent data set was derived by averaging results from the 36 analyses that used cover per cent data but varied other settings. The process was repeated for each of our 13 different data sets/settings: data set type ($n = 3$), SIMPER cumulative contribution score cut-off ($n = 2$) and user-defined SAAP settings ($n = 8$).

We calculated means and standard deviations for these 13 combinations of results, when the expert-assigned vegetation type was ranked first by the SAAP, and when it was ranked among the top five vegetation types by the SAAP. Analysis of variance (ANOVA) and *t*-tests were used to test for significant differences between means. We further explored these findings using histograms that showed the number of times the expert-assigned vegetation type was ranked first, second, third, to tenth by individual SAAP analyses.

An additional analysis tested how well the SAAP assigned plots with vegetation types based on the composition and cover of species at individual plots. Although we consider this a less robust approach than one based on the synoptic table analogue data derived from quantitative analysis, we anticipate the need for new plot data to be incorporated into an existing quantitatively derived or expert-defined vegetation classification.

Results

Eight hundred and sixty-seven floristic plots were manually assigned to 62 *a priori* VCA vegetation types, a process that required considerable expertise and took more than 70 h to complete. Sixty-three plots could not be manually assigned by the expert to an *a priori* VCA vegetation type and were labelled as belonging to a 'riverine grassland complex'. Trees were considered the dominant stratum for 42 VCA types containing 667 plots; shrubs for 10 types containing 140 plots; and the ground-cover for 10 types containing 60 plots (see Appendix S2). SIMPROF produced a variable number of statistically supported dendrogram groups depending on data set type (cover per cent, cover score, or presence-absence); with many fewer groups identified using the cover per cent data set type (Table 1).

Data set type was clearly an important determinant of concordance between SAAP results and the expert's results (Table 2). Significant differences among the mean number of

Table 1. SIMPROF results for the three data set types submitted for quantitative analysis of vegetation data

Data set type	Total number of SIMPROF identified dendrogram groups	Number of plots within groups	Number of single plot groups
Cover per cent (0–100)	58	1–155	5
Cover score (0–6)	95	1–64	10
Presence-absence (1, 0)	83	1–66	3

Table 2. The average number of plots for which the expert-assigned vegetation type was ranked highest, or among the highest five types, by grouped semi-automated assignment program (SAAP) analyses

Data set type	Ranked 1st		Ranked 1st–5th	
	Mean	Standard deviation	Mean	Standard deviation
Cover per cent	282 ^a	9.8	555 ^a	4.9
Cover score	200 ^b	2.3	510 ^b	3.6
Presence-absence	179 ^b	3.2	432 ^c	2.4
SIMPER characteristic species contribution score cut-off				
90%	246 ^a	10.0	516 ^a	8.0
99%	194 ^b	3.2	482 ^b	5.2
Number of characteristic species considered				
TOP 3	241	9.0	508	7.2
TOP 5	215	6.6	498	7.0
TOP 10	204	8.8	492	6.8
Down-weighting applied to lower ranked characteristic species				
1.00	209	10.1	493	7.1
0.75	222	6.6	508	6.5
0.50	229	6.6	495	7.3
Increased weighting of dominant stratum species				
1.00	216	8.0	487 ^b	7.0
1.25	224	7.9	511 ^a	6.9

All 108 SAAP analyses were repeatedly grouped by a single common data set type, or user-defined setting, to calculate means and standard deviations for that group. Superscripts denote significantly different means.

expert-assigned vegetation types ranked first by the SAAP ($F_{2,105} = 37.9$, $P < 0.001$), and ranked first to fifth by the SAAP ($F_{2,105} = 130.3$, $P < 0.001$), were reported between data set types. *Post-hoc* least-significant-difference tests revealed that for the first-ranked vegetation types, the cover per cent data set mean was significantly higher than the means for the other two data sets, which were not significantly different (Table 2). However, means based on the top five ranked vegetation types were significantly different for all three input data sets (Table 2). These results showed that reducing the influence of dominant species on the quantitative analyses (from cover per cent, to cover score, to presence-absence) significantly reduced the success of the SAAP in assigning plots to the same vegetation types as the expert. The results also showed that although the best result for types ranked first was only 282/867, or 33% of plots, when the first five SAAP ranked types

were considered, the success rate increased to an average of 555/867 or 64% of plots (Table 2).

Greater concordance resulted from the 90% SIMPER cumulative contribution score cut-off compared with the 99% cut-off (Table 2), and these results were significant, both for first place matches ($t_{106} = 4.3$, $P < 0.001$), and first to fifth place matches ($t_{106} = 3.1$, $P = 0.003$). Therefore, as more species were included in the synoptic table analogue data sets, concordance between SAAP and the expert's results decreased.

Three user-defined settings within the SAAP analyses were also tested for their influence on concordance. The first compared results when the number of *a priori* vegetation type characteristic species per stratum considered by the SAAP was varied from Top 3, to Top 5, to Top 10. Table 2 shows that the number of first, and first to fifth, ranked type matches decreased as the number of characteristic species considered increased. Although differences among means were not statistically significant, the results for types ranked first by the SAAP were close to significant ($F_{2,105} = 2.8$, $P = 0.066$). The second user-defined setting compared results when lower ranked characteristic species were down-weighted. ANOVA did not find significant differences between the three down-weighting settings for either first place, or first to fifth place results. However, for the final comparison based on increased weighting of dominant stratum species, significantly more matches were recorded among the first five SAAP ranked types when dominant stratum species received a weighting of 1.25 ($t_{106} = -2.14$, $P = 0.034$, Table 2).

Histograms based on the statistically significant cover per cent data set, SIMPER cumulative contribution score cut-off of 90%, and dominant stratum weighting of 1.25, were generated using results derived by varying the number of vegetation type characteristic species per stratum (Top 3, Top 5, Top 10), and down-weighting species according to their rank. The highest concordance was produced when only the Top 3 characteristic species per stratum were considered by the SAAP analyses (Fig. 2a–c). For this analysis, the expert-assigned vegetation type was ranked first by the SAAP analysis for 50% of plots (430/867, Fig. 2a). Levels of concordance were similar for the Top 3 and Top 5 characteristic species analyses when the first five SAAP ranked types were considered. In both cases the expert-assigned vegetation type was represented in the first five SAAP ranked types for 71% of plots (Fig. 2a,b). Concordance was considerably lower when the Top 10 characteristic species per stratum were considered (Fig. 2c). Down-weighting the contribution scores of the second and third characteristic species within each stratum did not improve concordance (Fig. 2a,d,e).

When plot-to-type assignment in the absence of quantitative analysis was explored, results again showed the trend of decreasing concordance with an increasing number of characteristic species considered by the SAAP (Table 3). The number of plots for which the expert-assigned vegetation type was ranked first by the SAAP, ranged from 298 (Top 3) to 227 (Top 10) plots, compared with 430 (Top 3) to 370 (Top 10) when the synoptic table analogue data were used (Fig. 2a–c). However in contrast to the results following quantitative analysis, the simple plot-to-type results showed a more gradual reduction in

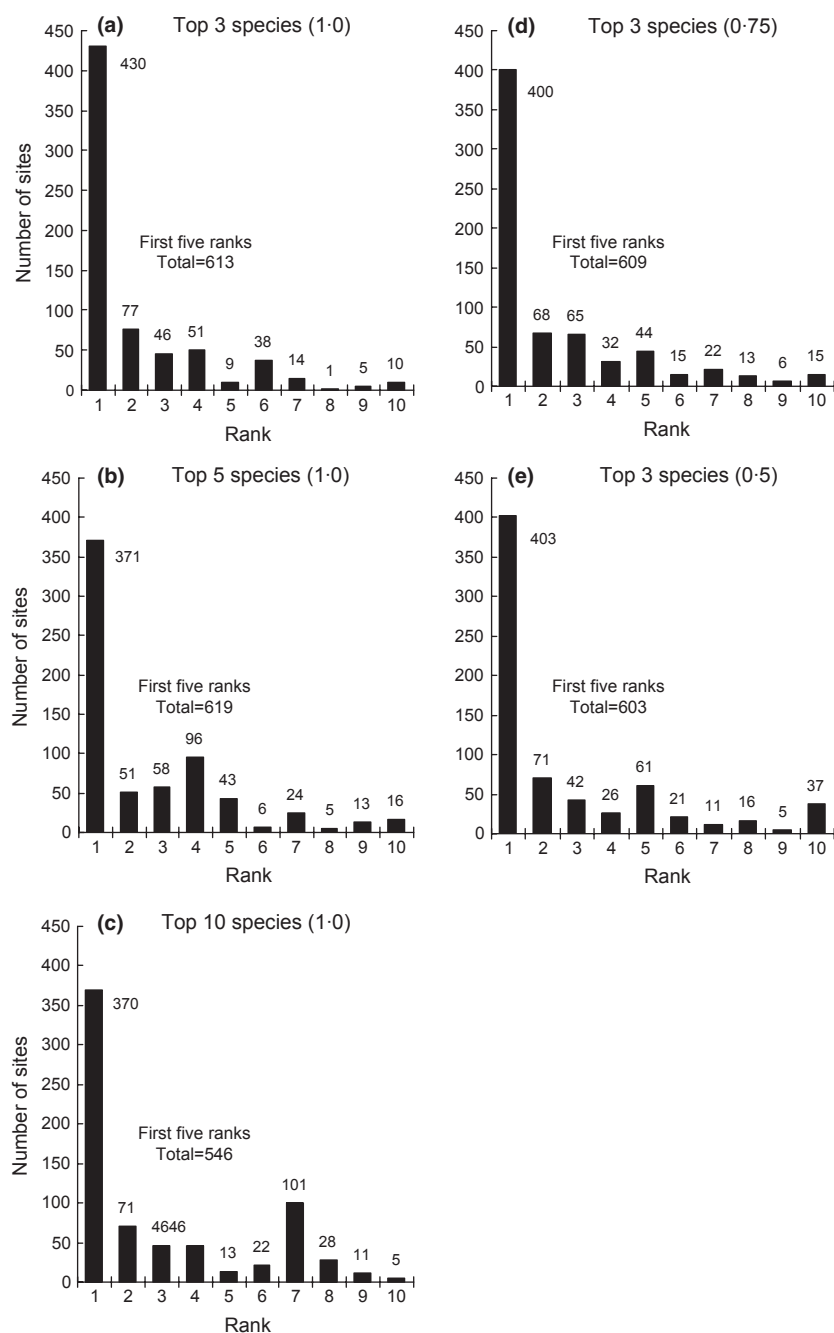


Fig. 2. The number of plots where the expert-assigned vegetation type was ranked first to tenth by individual semi-automated assignment program analyses that kept constant: cover per cent data, cumulative contribution cut-off score of 90%, and a dominant stratum weighting of 1.25 but varied the number of vegetation type characteristic species considered (a–c), and the weighting (shown in brackets) applied to the Top 3 species (a, d, e).

Table 3. Semi-automated assignment program results generated by simple plot-to-type assignment without prior quantitative analyses

Number of VCA characteristic species considered	Number ranked 1st	Number ranked 2nd	Number ranked 3rd	Number ranked 4th	Number ranked 5th	Total Number 1st–5th
Top 3	298	165	74	56	44	637
Top 5	258	144	91	62	46	601
Top 10	227	132	91	67	53	570

Input data and user selections are the same as for Fig. 2a,c.

VCA, Vegetation Classification and Assessment database.

the number of plots ranked first, then second, then third and so on (Table 3). This gradual decline resulted in a similar number of plots ranked within the top five SAAP ranked vegetation types for both the simple plot-to-type assignment process, and the more robust quantitative analysis based process.

Discussion

Many countries are investing in vegetation field survey and vegetation classification across regional to national scales (Grossman *et al.* 1998; Mucina *et al.* 2000; Chytrý, Schaminée & Schwabe 2011; Dengler *et al.* 2011; Wiser *et al.* 2011; De Cáceres & Wiser 2012), and large national vegetation databases continue to grow (Hennekens & Schaminée 2001; Schaminée *et al.* 2009, 2011), with more than 76 million vegetation plot data now available globally (GBIF Data Portal, <http://data.gbif.org/welcome.htm>). Using expert opinion to assign vegetation plots to vegetation types manually within existing classifications challenges an individual's mental resources. Even when ignoring the possibility that expert opinion can vary depending on a number of external factors (see Maule, Hockey & Bdzola 2000; Danziger, Levav & Avnaim-Pesso 2011) there is potential for the nature and quality of decision making to suffer given the time and effort required to manually allocate plots to an *a priori* classification. The manual process will also be constrained by the number of species an expert can consider simultaneously, which may be as few as four (see Cowan 2001).

This cognitive limitation associated with manual expert-dependent assignment processes has two important implications: Firstly, our study is based on a relatively small data set of 630 species from 930 plots, assigned to one of 99 *a priori* vegetation types. Therefore, larger studies with many more plots, species and vegetation types, are likely to prove intractable for a manual, expert-dependent assignment process. There is clearly a need for semi-automated processes to assist vegetation experts assign large numbers of vegetation plot data within pre-existing vegetation classifications (Gégout & Coudun 2012). Secondly, the expert-derived benchmark results, to which the results generated by our semi-automated process were compared, may not have represented the 'true' assignment to type for all plots. Despite the difficulty of ascertaining 'true' assignment due to vegetation being a continuously varying phenomenon (see Mucina 1997; French, Callaghan & Hill 2000; De Cáceres *et al.* 2009; Jennings *et al.* 2009; De Cáceres, Font & Oliva 2010; Chytrý, Schaminée & Schwabe 2011; De Cáceres & Wiser 2012), the results generated by our semi-automated process may in fact have been more accurate than those of the expert, a possibility also considered in similar studies by van Tongeren, Gremmen & Hennekens (2008) and Gégout & Coudun (2012). We therefore encourage vegetation scientists to further explore the effectiveness and efficiency gains provided by our process and the SAAP using their own independent vegetation data sets and classifications.

Our semi-automated process has been designed to assist with plot-to-type assignment within early-stage classifications based on lists of characteristic species. It begins with repeatable

and objective quantitative analysis of the available vegetation plot data, and develops synoptic table analogue data which are used by the SAAP to calculate goodness-of-fit scores between plots and *a priori* vegetation types. We consider this coupling of repeatable and objective quantitative analysis with the SAAP to be essential, especially in countries such as Australia, where early-stage classifications have been primarily based on expert knowledge. This is because quantitative analysis also provides the outputs required to evaluate how well these early-stage classifications do account for the full range of variation in vegetation composition and structure within regions (Mucina 1997; van Tongeren, Gremmen & Hennekens 2008; De Cáceres *et al.* 2009; Jennings *et al.* 2009; De Cáceres, Font & Oliva 2010).

In agreement with other authors (Rodwell 2006), we do not advocate the removal of the vegetation expert from the assignment process. The SAAP rapidly provides a transparent and defensible goodness-of-fit score between plots within dendrogram groups, and *a priori* vegetation types. Final plot-to-type assignment may be rapid and straight-forward when the first-ranked vegetation type receives a much higher goodness-of-fit score than the second. However, in some cases ranks may be tied, or scores for lower ranked vegetation types may decrease gradually. In these situations, other information such as location, landscape position, and soil type, are likely to aid final assignment. Some *a priori* vegetation types are floristically very similar and may share similar distributions, landscape positions and soil types [for example, see the River Red Gum (*Eucalyptus camaldulensis*) types present in our data set, Appendix S2]. SAAP's ranking of candidate types, and allocation of goodness-of-fit scores, will be invaluable for guiding expert review and final assignment in these situations. In addition, where final assignment is particularly problematic, the quantitative contributions of individual species (summed to provide the goodness-of-fit scores) may provide additional information to assist with final assignment (see Methods; Tables S2 and S3). In other cases, insufficient floristic information may be contained within the data to make a confident assignment. For example, plots may be located in vegetation types not recognised by the classification, or in disturbed vegetation dominated by widespread common species, or on ecotones (see Gégout & Coudun 2012 for an approach to identifying plots on ecotones). As currently exists for TABLEFIT (Hill 1989, 1996), we expect that further application of the SAAP to new data sets will lead to recommendations on goodness-of-fit cut-off values below which groups of plots cannot be confidently assigned.

Our study found clear evidence supporting the use of particular data set types, quantitative analyses, and SAAP user-defined settings, for maximising the concordance of results with those of the expert. Significantly higher concordance was achieved using the 0–100% cover per cent data set. This suggests that dominance expressed in raw cover terms was the primary driver behind the manual expert assignment process. Whether this data set type (raw cover per cent) will be appropriate for real world applications will depend on whether the target classification's vegetation types were similarly identified and constructed based on raw cover data. Concor-

dance was also higher when our synoptic table analogue data were based on SIMPER cumulative contribution cut-off scores of 90% rather than 99%. That is, concordance was greater when fewer species were included in the synoptic table analogues. This may be due to the expression of more stochastic variation, in species composition and structure within synoptic table analogues when more species are included (Jennings *et al.* 2009).

We similarly found better concordance when fewer *a priori* vegetation type characteristic species were considered by the SAAP. Further knowledge of the *a priori* classification used in this study may help to explain this result. From a total of 99 vegetation types, 72 were defined at the classification level of association, and within Australia, associations tend to be defined using the three most dominant species within each stratum (see Australian National Vegetation Information System in ESCAVI 2003). We may therefore have anticipated that our highest concordance for these types would be achieved when only the Top 3 species within each stratum were considered by the SAAP. Understanding the conceptual and practical underpinnings of the *a priori* vegetation types is therefore an important prerequisite to selecting the most appropriate data set type (e.g. cover per cent, cover score, presence-absence, or an alternative) to submit to quantitative analysis, as well as selecting the number of characteristic species per stratum to be considered by the SAAP. This aligns with the need for consistency in assignment of new vegetation observations to previously defined vegetation types, in accordance with how the types were originally defined (De Cáceres & Wiser 2012).

Our study demonstrated that when the first five ranked vegetation types were considered, there was little difference in concordance between the expert's and SAAP's results based on the Top 3 or Top 5 characteristic species per stratum. Consequently, our recommendation for application of the SAAP to future studies is to undertake SAAP analyses based on the Top 5 characteristic species per stratum and expertly review up to the first five ranked vegetation types that result. Analyses based on the Top 5 characteristic species per stratum will be essential when types are defined at the level of sub-association ($n = 13$ in this study, see ESCAVI 2003), and also when vegetation types characterised by only one or two major strata are part of the classification. In addition, and especially where grassland types are being considered, our findings also support the use of an increased weighting being applied to species within the dominant stratum (1.25 used here). Applying the SAAP with these settings to our 63 'riverine grassland complex' plots that could not be manually assigned to an *a priori* vegetation type by our expert, would instantly generate a ranked list of candidate vegetation types each supported by its goodness-of-fit cumulative contribution score.

Vegetation classification, modelling, mapping and monitoring are undergoing rapid evolution and becoming more computer intensive and dependent on remotely sensed data. However, the accuracy of the products developed will be closely related to the appropriate analysis of available vegetation plot data. An important aim of such analyses will be the effective and efficient assignment of plots within an *a priori* vegetation classification. Our semi-automated process,

designed for early-stage classifications, supports this aim. It combines repeatable and objective quantitative analyses, with the SAAP, to calculate quantitative goodness-of-fit scores between vegetation plots and types, based on the species that characterise each *a priori* vegetation type, and the species that characterise groups of plots derived from quantitative analyses. Importantly, our semi-automated process also promotes the on-going evaluation of early-stage vegetation classifications, by developing the numerical classification outputs upon which these evaluations can be based (Mucina 1997; De Cáceres *et al.* 2009; Jennings *et al.* 2009; De Cáceres, Font & Oliva 2010; JNCC 2011).

Acknowledgements

Thanks are extended to Ken Turner for data export and nomenclatural review and Stephan Hennekens for introducing us to the program ASSOCIA. The manuscript was improved based on comments received from Sarah Hill, three anonymous reviewers, and the associate editor of this journal.

References

- Anderson, M., Bourgeron, P., Bryer, M.T., Crawford, R., Engelking, L., Faber-Langendoen, D., Gallyoun, M., Goodin, K., Grossman, D.H., Landaal, S., Metzler, K., Patterson, K.D., Pyne, M., Reid, M., Sneddon, L. & Weakley, A.S. (1998) *International Classification of Ecological Communities: Terrestrial Vegetation of the United States. Volume II. The National Vegetation Classification System: List of Types*. The Nature Conservancy, Arlington, Virginia.
- Benson, J.S. (2006) New South Wales Vegetation Classification and Assessment: introduction – the classification, database, assessment of protected areas and threat status of plant communities. *Cunninghamia*, **9**, 331–382.
- Benson, J.S., Richards, P.G., Waller, S. & Allen, C.B. (2010) New South Wales Vegetation Classification and Assessment: part 3 Plant communities of the NSW Brigalow Belt South, Nandewar and west New England Bioregions and update of NSW Western Plains and South-western Slopes plant communities, Version 3 of the NSW VCA database. *Cunninghamia*, **11**, 409–579.
- Chytrý, M., Schaminée, J.H.J. & Schwabe, A. (2011) Vegetation survey: a new focus for Applied Vegetation Science. *Applied Vegetation Science*, **14**, 435–439.
- Clarke, K.R. & Gorley, R.N. (2006) *PRIMER V6*. PRIMER-E Ltd, Plymouth.
- Clarke, K.R., Somerfield, P.J. & Gorley, R.N. (2008) Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. *Journal of Experimental Marine Biology and Ecology*, **366**, 56–69.
- Cowan, N. (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioural and Brain Sciences*, **24**, 87–114.
- Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, **108**, 6889–6892.
- De Cáceres, M., Font, X. & Oliva, F. (2010) The management of vegetation classifications with fuzzy clustering. *Journal of Vegetation Science*, **21**, 1138–1151.
- De Cáceres, M. & Wiser, S.K. (2012) Towards consistency in vegetation classification. *Journal of Vegetation Science*, **23**, 387–393.
- De Cáceres, M., Font, X., Vicente, P. & Oliva, F. (2009) Numerical reproduction of traditional classifications and automatic vegetation identification. *Journal of Vegetation Science*, **20**, 620–628.
- Dengler, J., Jansen, F., Glöckler, F., De Cáceres, M., Chytrý, M., Ewald, J., Oldeland, J., Lopez-Gonzalez, G., Finckh, M., Mucina, L., Rodwell, J., Schaminée, J. & Spencer, N. (2011) The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science. *Journal of Vegetation Science*, **22**, 582–597.
- ESCAVI (Executive Steering Committee for Australian Vegetation Information) (2003) *National Vegetation Information System – Australian Vegetation Attribute Manual Version 6.0*. Department of Environment and Heritage, Canberra.
- French, K., Callaghan, B. & Hill, S. (2000) Classifying endangered vegetation communities: a case-study of Cumberland Plain Woodlands. *Pacific Conservation Biology*, **6**, 120–129.
- Gégout, J. & Coudun, C. (2012) The right relevé in the right vegetation unit: a new typicality index to reproduce expert judgement with an automatic classification programme. *Journal of Vegetation Science*, **23**, 24–32.
- Grossman, D.H., Faber-Langendoen, D., Weakley, A.S., Anderson, M., Bourgeron, P., Crawford, R., Goodin, K., Landaal, S., Metzler, K., Patterson,

- K., Pyne, M., Reid, M. & Sneddon, L. (1998) *International Classification of Ecological Communities: Terrestrial Vegetation of the United States. Volume 1. The National Vegetation Classification System: Development, Status, and Applications*. The Nature Conservancy, Arlington, Virginia.
- Hennekens, S.M. & Schaminée, J.H.J. (2001) Turboveg, a comprehensive database management system for vegetation data. *Journal of Vegetation Science*, **12**, 589–591.
- Hill, M.O. (1989) Computerized matching of relevés and association tables, with an application to the British National Vegetation Classification. *Vegetatio*, **83**, 187–194.
- Hill, M.O. (1996) *TABLEFIT Version 1.0 for Identification of Vegetation Types*. Institute of Terrestrial Ecology, Huntingdon. <http://www.ceh.ac.uk/products/software/index.html> [accessed 29 February 2012].
- Jennings, M.D., Faber-Langendoen, D., Louckes, O.L., Peet, R.K. & Roberts, D. (2009) Standards for associations and alliances of the U.S. National Vegetation Classification. *Ecological Monographs*, **79**, 173–199.
- JNCC (2011) A compilation of proposed additions and revisions to vegetation types in the National Vegetation Classification. Joint Nature Conservation Committee. <http://jncc.defra.gov.uk/page-4269> [accessed 29 February 2012].
- Kočí, M., Chytrý, M. & Tichý, L. (2003) Formalized reproduction of an expert-based phytosociological classification: a case-study of subalpine tall-forb vegetation. *Journal of Vegetation Science*, **14**, 601–610.
- Maule, A.J., Hockey, G.R.J. & Bdzola, L. (2000) Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica*, **104**, 283–301.
- McCune, B. & Grace, J.B. (2002) *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, Oregon.
- Mucina, L. (1997) Classification of vegetation: past, present and future. *Journal of Vegetation Science*, **8**, 751–760.
- Mucina, L., Bredenkamp, G.J., Hoare, D.D. & McDonald, D.J. (2000) A national vegetation database for South Africa. *South African Journal of Science*, **96**, 497–498.
- Rodwell, J.S. (2006) *National Vegetation Classification: Users' Handbook*. Joint Nature Conservation Committee, Peterborough.
- Sattler, P.S. & Williams, R.D. (eds) (1999) *The Conservation Status of Queensland Bioregional Ecosystems*. Environmental Protection Agency, Brisbane.
- Schaminée, J.H.J., Hennekens, S.M., Chytrý, M. & Rodwell, J.S. (2009) Vegetation-plot data and databases in Europe: an overview. *Preslia*, **81**, 173–185.
- Schaminée, J.H.J., Janssen, J.A.M., Hennekens, S.M. & Ozinga, W.A. (2011) Large vegetation databases and information systems: new instruments for ecological research, nature conservation, and policy making. *Plant Biosystems – An International Journal Dealing with all Aspects of Plant Biology*, **145**, 85–90.
- Tichý, L. (2002) JUICE, software for vegetation classification. *Journal of Vegetation Science*, **13**, 451–453.
- Tichý, L., Chytrý, M., Hájek, M., Talbot, S.S. & Botta-Dukát, Z. (2010) Optim-Class: using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science*, **21**, 287–299.
- van Tongeren, O., Gremmen, N. & Hennekens, S. (2008) Assignment of relevés to pre-defined classes by supervised clustering of plant communities using a new composite index. *Journal of Vegetation Science*, **19**, 525–536.
- Willner, W. (2011) Unambiguous assignment of relevés to vegetation units: the example of the *Festuco-Bometa* and *Trifolio-Geranieta sanguinei*. *Tuexenia*, **31**, 271–282.
- Wiser, S.K., Hurst, J.M., Wright, E.F. & Allen, R.B. (2011) New Zealand's forest and shrubland communities: a quantitative classification based on a nationally representative plot network. *Applied Vegetation Science*, **14**, 506–523.

Received 17 May 2012; accepted 3 September 2012

Handling Editor: Darren Kriticos

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Excerpt showing the header row and first few records of the characteristic species SAAP input file prepared using information contained within the NSW Vegetation Classification and Assessment (VCA) database.

Table S2. Characteristic species input file from SIMPER analysis of one SIMPROF identified dendrogram group.

Table S3. Excerpt from full matrix output for one branch group generated by SAAP analyses.

Table S4. Excerpt from summary matrix output for eight branch groups generated by SAAP analyses showing the top five ranked VCA vegetation types for each of eight dendrogram groups.

Appendix S1. Survey details for the 930 plots submitted to SAAP analysis.

Appendix S2. NSW Vegetation Classification and Assessment (VCA) types assigned using expert opinion to 867 full-floristic plots within the Murray CMA region.

Dataset. Features of the case-study data set and the semi-automated assignment program (SAAP).

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.