To appear as Chapter 4 in J. Franklin and E. van der Maarel, eds. 2012. *Vegetation Ecology*. Second edition. Oxford University Press, New York, NY.

Classification of natural and semi-natural vegetation

Robert K. Peet and David W. Roberts

4.1 Introduction

Vegetation classification has been an active field of scientific research since well before the origin of the word ecology and has remained so through to the present day. As with any field active for such a long period, the conceptual underpinnings as well as the methods employed, the products generated and the applications expected have evolved considerably. Our goal in this chapter is to provide an introductory guide to participation in the modern vegetation classification enterprise, as well as suggestions on how to use and interpret modern vegetation classifications. The historical development of classification and the associated evolution of community concepts is provided by van der Maarel & Franklin in Chapter 1, and Austin describes numerical methods for community analysis in Chapter 2. While we present some historical and conceptual context, our goal in this chapter is to help the reader learn how create, interpret and use modern vegetation classifications, particularly those based on large-scale surveys.

4.1.1 Why classify?

Early vegetation classification efforts were driven largely by a desire to understand the natural diversity of vegetation and the factors that create and sustain it. Vegetation classification is critical to basic scientific research as a tool for organizing and interpreting information and placing that information in context. To conduct or publish ecological research without reference to the type of community the work was conducted in is very much like depositing a specimen in a museum without providing a label. Documentation of ecological context can range from a simple determination of the local community context to a detailed map showing a complicated spatial arrangement of vegetation types as mapping units. This need for documenting ecological context is also scale transgressive with vegetation classification schemes contributing equally to research from small populations of rare species to that involving global projection of human

impacts (Jennings *et al.* 2009). Frameworks other than vegetation classification are conceivable for documenting ecological context. For example, environmental gradients and soil classifications have often been used to define site conditions. However, these require *a priori* knowledge of factors important at a site while vegetation classification, in contrast, lets the assemblage of species and their importance serve as a bioassay.

Use of vegetation classification has increased over the last few decades. Vegetation description and classification provides units critical for inventory and monitoring of natural communities, planning and managing conservation programs, documenting the requirements of individual species, monitoring the use of natural resources such as forest and range lands, and providing targets for restoration. Vegetation types are even achieving legal status where they are used to define endangered habitats and where protection of these types is mandated such as with types deemed in need of protection under European law (see Waterton 2002). For example, the European Union has created lists of protected vegetation types, and vegetation types are being used to develop global red lists of threatened ecosystems (*e.g.* Rodríguez *et al.* 2011).

4.1.2 The Challenge

The goal of vegetation classification is to identify, describe and inter-relate relatively discrete, homogeneous, and recurrent assemblages of co-occurring plant species. Vegetation presents special challenges to classification as it varies more or less continuously along environmental gradients and exhibits patterns that result from historical contingencies and chance events (Gleason 1926, 1939). Not surprisingly, multiple solutions are possible and as Mucina (1997) and Ewald (2003) have explained, adopting one approach over another should be based on practical considerations.

Although there is considerable variability in approaches taken to vegetation classification, most initiatives embrace some basic assumptions about vegetation and its classification. Four such widely adopted assumptions were articulated by Mueller-Dombois & Ellenberg in their classic 1974 textbook. (1) Similar combinations of species recur from stand to stand under similar habitat conditions, though similarity declines with geographic distance. (2) No two stands (or sampling units) are exactly alike, owing to chance events of dispersal, disturbance, extinction, and history. (3) Species assemblages change more or less continuously if one samples a geographically widespread community throughout its range. (4) Stand similarity varies with the spatial and temporal scale of analysis. These underlying assumptions have led to the wide adoption of a practical approach wherein community types are characterized by attributes of vegetation records that document similar plant composition and physiognomy with the vegetation classification relying on representative field records (plots) to define the central concept of the type. Subsequent observations of vegetation are determined as belonging to a unit through their similarity to the type records for the individual communities.

Another challenge that increasingly confronts the vegetation classification enterprise is that with the widespread adoption of classification systems for inventory, monitoring,

management and even legal status, classification systems need to have comprehensive coverage, stability in the classification units, plus a transparent process for revising those units. This new and broader set of applications suggests that we need to move toward consensus classifications that combine the inquiry of many persons into a unified whole, and that the rules for participation be open and well defined. However, we must also recognize that as the applications of vegetation classification migrate from the pure scientific arena to one of management and policy, the categories are likely to evolve in ways that find their origin not just in science but also in policy and public opinion (Waterton 2002).

4.2 Classification frameworks: history and function

Vegetation classifications systems can vary from local to global and from fine-scale to coarse-scale, and the approaches to vegetation classification employed tend to reflect the scale of the initiative. Classification schemes used at the global scale tend to focus on growth forms or physiognomic types that reflect broad-scale climatic variation rather than species composition (discussed by Box & Fujiwara in Chapter 15). In the present chapter our focus is on actual or realized natural and semi-natural vegetation. These are generally bottom-up classifications where units are defined by sets of field observations where species occurrences and/or abundances were recorded. Vegetation classification has a rich history (discussed in Chapter 1). Whittaker (1962, 1973), Shimwell (1971) and Mueller-Dombois & Ellenberg (1974) all review the many and varied approaches that have been applied to vegetation classification. Subsequent synthetic overviews by Kent (2012), McCune & Grace (2002), and Wildi (2010) summarize, compare and evaluate commonly used methods.

Although local-scale projects can use any classification criteria that provide a convenient conceptual framework for the project at hand, such local and idiosyncratic classifications do not allow the work to be readily placed in a larger context. The growing recognition of the need for vegetation classification research to place new results in context means that a consistent conceptual framework is needed for all components of the classification process (De Cáceres & Wiser 2012). Below we summarize key components of two such frameworks: European phytosociology as it has evolved from the school of Braun-Blanquet, and the much more recently developed U.S. National Vegetation Classification. We then summarize the differences and compare these classifications to those encountered in other national-level initiatives.

4.2.1 The Braun-Blanquet approach and contemporary European phytosociology

By far the most widely applied approach to vegetation classification is that advocated by Josias Braun-Blanquet. The method centers on recording fine-scale vegetation composition. The basic unit of observation is the plot (or relevé) within which all species are recorded by vertical stratum and the abundance of each is estimated, usually employing an index of cover/abundance. Related plots are combined in tabular form and groups of similar plots are defined as communities based on consistency of composition. The basic unit, adopted at the

International Botanical Congress in 1910, is the association, which is defined as having “definite floristic composition, presenting a uniform physiognomy, and growing in uniform habitat conditions.” The community is then characterized by the constancy of shared taxa and specific diagnostic species that provide coherency to the group and set it off from other groups. Historically, table sorting was done by hand, while today computer-aided sorting is the rule with numerous algorithms available to automate the process (see section 4.6.1 for more detail, or consult Braun-Blanquet 1964 or Westhoff & van der Maarel 1973). Similar associations that share particular diagnostic species are combined into higher level assemblages, there being five primary levels (Association, Alliance, Order, Class, and Formation).

Once an author has developed one or more new or revised associations, that author reviews past published work, designates the critical diagnostic species, assigns a unique name following the International Code of Phytosociological Nomenclature (Weber *et al.* 2000), places it within the hierarchy and submits the work for publication. The process is similar to that required to establish a new species. In both cases the author examines documented occurrences, writes a monograph wherein the examined occurrences are typically reported, and specifies plots or a type specimen that serve to define the type. In the Braun-Blanquet system one plot is designated the nomenclatural type for each association, the nomenclature follows a formal code that gives priority to the first use of a name, and the resultant associations are then available in the literature for scientists to discover and accept or not.

The strongest attributes of the Braun-Blanquet system are the consistency of the approach, the enormous number of plots that have been recorded (with an estimated total for Europe alone of 4.3 million; Schaminée *et al.* 2009), and the large number of published vegetation descriptions of types. Weaknesses include a seeming arbitrary definition of units, the lack of requirement that new units be integrated with established units, and the lack of any formal registry of published units. Some potential users find the naming system awkward, which is why the recent vegetation classification of Great Britain divorced itself from the traditional nomenclature, despite the fundamental units otherwise closely approximating the associations of the Braun-Blanquet system (Waterton 2002, Rodwell 2006).

The literature on European vegetation is so enormous that summarizing it has proven extremely difficult. Community types have been synthesized for quite a few countries and other geographic units, but these efforts have not yet been integrated. In 1992 The European Vegetation survey was established with the goal of fostering collaboration and synthesis (Mucina *et al.* 1993, Rodwell *et al.* 1995). One direct result has been a number of trans-national overviews of thematic types and a summary of types at the alliance level and above by Rodwell *et al.* (2002). In addition, there has been movement toward standards for collecting plot data (Mucina *et al.* 2000), and the development of the software program TurboVeg (Hennekens & Schaminée 2001) for managing plot data has led to considerable standardization in data content and format.

4.2.2 The United States National Vegetation Classification

The development of the US National Vegetation Classification (USNVC) provides clear contrasts with the European classification enterprise, although both have roughly equivalent primary units (in both cases called associations), and both are based on vegetation plot records. Historically, when vegetation classification was undertaken by North America by academic ecologists, the approaches tended to be idiosyncratic and specific to the particular project. In the absence of leadership from the academic community, various federal land management and environmental regulatory agencies in the U.S. created classifications systems for their own purposes, such as for wetlands (Cowardin *et al.* 1979), land-cover (Bailey 1976), and forest vegetation (Pfister & Arno 1980).

Vegetation classification in the United States has matured considerably over the past two decades in response to three initiatives. First, starting in the 1970s, The Nature Conservancy, a non-governmental organization, encouraged the development of state programs to inventory the status of biodiversity for conservation planning. The lack of consistency in inventory units between states ultimately led to a national vegetation classification system based on types provided by state programs, published literature, and expert opinion (Anderson *et al.* 1998). Although at first largely subjective, the units were defined to be non-overlapping and to constitute a formal list of recognized types. This effort led to an international classification (see Grossman *et al*. 1998, Anderson *et al*. 1998, Jennings *et al.* 2009). As this system has matured, emphasis has been placed on both providing linkage to original data and describing the variation in each type across its geographic range. Second, growing recognition of the need for common standards for geospatial data across government agencies led to the establishment of the US Federal Geographic Data Committee (USFGDC), including a subcommittee for standardizing vegetation classification activities across government agencies. Although this standard is formally recognized only for cross-tabulating classifications, it is beginning to have broad application in its own right. Third, members of the Ecological Society of America (ESA) recognized the diversity of approaches and standards in use across the country, the need to allow broad participation by interested parties, and the importance of peer review of proposed changes in the classification. ESA established a Panel on Vegetation Classification in 1994 that subsequently proposed standards for vegetation classification (Jennings *et al.* 2009).

These three independent initiatives formed a formal partnership to advance the USNVC that led to adoption of a new USFGDC standard in 2008, including rules for documentation and peer review of proposed new and revised types. As a consequence of this partnership, the US has a national classification with a definitive set of associations (~6200 at this writing) with developing mechanisms for modification of this list by interested parties. By requiring that accepted types span the known range of variation, that they not overlap, and that they be based on vegetation records in public archives, the system is more forward looking than the European initiatives. However, at this time the US community types have only limited linkage to archived data, and the formal descriptions of types are not always described in sufficient detail to provide keys or expert systems for determining vegetation occurrences for most areas. Thus, while the US infrastructure is very progressive, the content will require considerably more development to catch up with the established European initiatives.

The USNVC formal hierarchy differs from that of the Braun-Blanquet system in that it is not derived entirely from lumping smaller units into larger ones. Instead it has three upper levels that provide a top-down, physiognomic hierarchy with units that are global in conception (Formation Class, Formation subclass and Formation). Nested below these are three middle levels based on biogeographic and regional environmental factors (Division, Macrogroup, and Group). At the base are Associations, which are combined into Alliances that nest into the middle-level Groups. This three-tier, eight-level hierarchy is intended to provide interpretable and widely applicable units across all spatial scales. The nesting is not always as seamless as it is in the Braun-Blanquet approach, but is intended to facilitate a broader range of applications.

4.2.3 Attributes of successful classification systems

The recent British National Vegetation Classification (NVC) program is a model for standardized data collection in a vegetation classification project and system. This program was led by John Rodwell who described the methodology in a user's handbook (Rodwell 2006). There are standard rules for placement of plots, size of plots and data to be collected. Standard forms were used to minimize drift in field methods. Any large new initiative would be well advised to adopt the level of standardization employed in the UK NVC, and small programs should adopt methods and goals consistent with well-established programs in order to maximize compatibility. The Braun-Blanquet, British and US initiatives all have their own standard nomenclatures, although the formats and the rules vary considerably between the systems. Finally, the US system remains unique in requiring public archiving of supporting plot data and providing systems for interested stakeholders to formally propose changes, both of which are likely to lead to more rigorously defined types.

4.3 Components of vegetation classification

There are ten primary components to vegetation classification, the complexity of the individual components depending on the situation, but all of them being important. We define those components here, and starting in section 4.4 we address each in some detail.

Project planning. Defining the geographic and ecological extent or range of the study allows data needs to be defined and existing data to be identified and evaluated. Often this will involve extensive preliminary work to aid in selection of field sites, perhaps through stratification relative to composition or environment or successional development, or in more human-dominated systems through searching out the remaining examples of natural and semi-natural vegetation.

Data acquisition. Once the objective of the study is defined, quantitative data characterizing vegetation composition must be acquired as new records or from databases of previously collected vegetation records. At a minimum, each record will contain the date and location of

observation, some attributes of the site, a list of plant taxa and some measure of importance for each taxon.

Data preparation. Before the vegetation composition data can be analyzed, the observation records need to be combined into a single dataset wherein inconsistences in field methods, scales of observation, measures of abundance, units of environment, resolution of species identifications and inconsistent taxonomic authorities have all been resolved. Although the goal is straight forward, complete integration without loss of information is often impossible and this component often involves a number of difficult and often subjective decisions.

Community entitation. This is the most essential step in classification as it is the recognition or creation of the entities that constitute the classification units. A broad range of methods can be employed, often iteratively and in combination, to define the classification units or "types." As vegetation often varies continuously in time and space, there is nothing conceptually as solid as a species and different investigators following different rules and protocols often come up with different classification units.

Cluster assessment. Once entities have been defined, it is important to critically analyze the results to determine that the types are relatively homogeneous and distinct from other types (Lepš & Šmilauer 2003), and to assure that distributions of species within types exhibit high fidelity and ecologically interpretable patterns. The criteria often involve formal assessment of the quantitative similarity (or dissimilarity) of vegetation composition within versus between types and the calculation of quantitative indices of species fidelity to types.

Community characterization. Entities must be characterized in a way that allows additional occurrences to be recognized with less than a full-scale reanalysis, and also allows placement in a larger system of community types. Traditionally, this has included assessment of the typical abundance and frequency of taxa, and in many cases identification of indicator species and the typical range of environmental conditions.

Community determination. Users of classifications need to be able to determine to which vegetation classification unit an instance of vegetation should be assigned, be it a published or archived record of vegetation or a new field observation. Tools range from dichotomous keys, to methods that use mathematical similarity, to expert systems. Determinations range from binary (yes/no), to multiple types with various designated degrees of fit.

Classification integration. Vegetation classification is often intended to expand or revise an established, large-scale vegetation classification system. Often this involves changes in established units, or replacement of previously published units. This, in turn, requires that levels of resolution (*e.g.*, fineness of splitting), criteria for peer review, and the importance of stability in classification systems be addressed systematically, more so than has historically been the case. For effective communication, community types need names, and the names need to be compliant with the current standards of the classification system (*e.g.* Weber et. al. 2000, USFGDC 2008).

Classification documentation. The results of vegetation classification initiatives need to be documented, both as to the units recognized and the data analyzed. Different classification systems have different requirements, formats and protocols. Publication with tables summarizing composition is always important, and vegetation records used in the analysis should be deposited in a public database.

4.4 Project planning and data acquisition

The fundamental unit for recording vegetation is the plot (or relevé). Associated with the plot are records of its location, size, physical setting and vegetation composition. The distribution and placement of plots, their size and shape, and the attributes to be recorded vary among recognized protocols and are important decisions to make when initiating a new project, or to recognize when using existing plot data.

4.4.1 Plot distribution and location

The first step in a vegetation classification project is definition of the geographic and compositional variation in vegetation to be classified as this will determine the number of plots needed and the difficulty of acquiring them. This step can be accomplished by literature review, conversation with regional experts, and preliminary field reconnaissance. Next, existing relevant vegetation plot data should be identified. This is not always straight-forward for while some plot data are available in public archives (*e.g.* VegBank; see www.vegbank.org, Peet *et al*. 2012), and many datasets are described in indices of plot databases (*e.g.* GIVD; see www.givd.info, Dengler *et al.* 2011), many datasets are not widely known and must be discovered by contacting likely sources. Once the availability of extant plot data is assessed and the need for new plots has been ascertained, the next step is to estimate the effort required to obtain those new plots.

The physical distribution of plots across the study area can be determined in a number of ways and these will reflect the objectives of the project. Traditionally, plots have been placed using preferential sampling where the investigator subjectively locates them to cover the range of variation needed for the project. The potential for bias in this method is obvious, so sometimes field plots are randomly located, or the landscape is stratified and plots are placed randomly within the strata. An alternative form of stratification often employed is the gradsect method where vegetation samples are stratified along known gradients of compositional variation (see Gillison & Brewer 1985, Austin & Heyligers 1989, 1992). As random and stratified sampling might under sample rare or unknown types, it is not uncommon for a probability designed sample to be supplemented with preferential plots on types poorly represented in the sample. Also, as the spatial extent of the project increases, the need for both stratification and some component of preferential sampling increases. For example, if one were sampling the range of variation in riparian vegetation across a moderate-sized European country or American state, there would inevitably be preferential selection of regions within which the sampling

would occur. In contrast, if the objective of the project were an inventory of the area of each vegetation type, or of standing timber, objective sampling methods would be more critical. An example of this is the Forest Inventory and Analysis plot system of the United States Forest Service designed to monitor the timber supply of the nation. This system uses a base grid of sample points with one plot located randomly in each of the 125,100 2430 ha hexagonal cells (Bechtold & Patterson 2004, Gray *et al.* 2012).

The potential for bias in preferentially located plots has led to considerable introspection and some critical analysis. Preferential sampling is often favored in human-manipulated landscapes where patches of natural and semi-natural vegetation tend to be small and influenced by recent land use. Roleček *et al.* (2007) explain that while probability designed sampling schemes better meet certain statistical assumptions, preferential sampling yields data sets that cover a broader range of vegetation variability including rare types that might otherwise have been missed. Random sampling is required when the sample units must represent a single statistical population. In vegetation sampling generally the intention is to distinguish types that are not necessarily members of the same statistical population, but rather are distinguishable entities.

Michalcová *et al.* (2011) further considered the problems inherent in using large plot databases wherein many of the plots are likely to represent preferential sampling. They found that sets of preferential samples contained more endangered species and had higher beta diversity, whereas estimates of alpha diversity and representation of alien species were not consistently different between preferentially and stratified-randomly sampled data. Thus, if the goal is to characterize the range of compositional variation or maximize species coverage, then at least some element of preferential sampling can be important.

4.4.2 Plot size and shape

Choice of plot size and shape can significantly influence perception of vegetation for a number of reasons. First, vegetation is spatially variable at nearly all scales. This variation can be driven by underlying environmental variation, biological interactions, or historical events (Nekola & White 1999). Secondly, species number increases with plot size; the logarithm of species richness usually varying directly with the logarithm of plot area (Fridley *et al.* 2005). The consequence is that as plot size increases, more species are encountered, as is more within-plot spatial variation. Plot shape has a similar tradeoff in that plots with low perimeter to volume ratios (squares and circles) tend to minimize spatial pattern (within-plot heterogeneity) and thus species number, whereas plots with high edge-volume ratios (*e.g.* long, thin plots) maximize representation of both patch types and species.

Historically, the solution to the tradeoff between homogeneity and completeness was to create a species-area curve to assess the “minimum area” needed to represent a particular type of vegetation. Unfortunately there is no objective stopping rule for plot area. In addition, plots were preferentially located in homogenous vegetation, but again this was subjective as some pattern can nearly always be found within a plot. Plot size also traditionally varied with

vegetation height so as to capture a snapshot of the total community, and Dengler *et al.* (2008) observe that as a rule of thumb plots are roughly as large in square meters as vegetation is high in decimeters.

In excess of four million vegetation plots are available in various archives (see Schaminée *et al.* 2009, Dengler *et al.* 2011). Integrating subsets of these plots for various analyses is complicated by the diversity of plot sizes and shapes. In addition, metrics such as species constancy and plot similarity can vary with plot size (Dengler *et al.* 2009). Collectively, these considerations have led a series of authors to propose that a standard set of plot sizes be adopted to facilitate future data integration and analysis. For example, Chytrý & Otýpková (2003) proposed plot sizes of 4 m^2 for sampling aquatic vegetation and low-grown herbaceous vegetation, 16 m^2 for grassland, heathland and other herbaceous or low-scrub vegetation types, 50 m^2 for scrub, and 200 m^2 for woodlands. In contrast, many North American ecologists have followed a tradition established by Whittaker (1960) of recording forest vegetation in 1000 m^2 plots, reflecting the generally higher tree species richness of North American forests as compared to European forests.

Peet *et al.* (1998) proposed that because there is no one correct scale for observing vegetation and that because different factors influence composition at different scales, vegetation should be recorded at multiple scales, both to facilitate data integration across projects and to allow investigation of processes working at different scales. They proposed a specific protocol with plots on a nearly log scale of 0.01, 0.1, 1, 10, 100, 400 and 1000 m^2. For their study they suggested 100 m^2 as the smallest acceptable total plot size, calling smaller-scale pattern 'within-community variation'. Such nested designs largely originated with Whittaker *et al.* (1979), with alternative protocols subsequently proposed (*e.g.* Stohlgren *et al.* 1995, Dengler 2009). All these protocols note increased variance in composition among subsamples at smaller scales and recommend that there be multiple small plots within each large plot to increase the range of this variance documented, the particular design of the nesting varying with the protocol. Although there is no consensus as to the optimal size or arrangement of nested plots, some form of nested sampling is highly desirable, if for no other reason than to maximize the potential for aggregating the plots with those from other studies. Moreover, with careful plot design, relatively little extra effort is required to include nested plots within the largest plot. Users of nested protocols should, however, be cautious not to aggregate dispersed subplots into larger subsamples as this will inflate species numbers owing to the subplots spanning an artificially high range of within community variation.

4.4.3 Plot records

In its simplest form, a plot record contains information about the observation event, the site, and the plants observed at the site. Lists of required and recommended plot attributes have been codified for numerous plot protocols and with remarkably similar prescriptions (*e.g.,* Mucina *et al.* 2000, Rodwell 2006, Jennings *et al.* 2009). These prescriptions often recognize two kinds of plots; occurrence plots are those used to determine the vegetation type at a site or document its presence, whereas classification plots are those intended for development or

improvement of a classification. Occurrence plots require only a subset of the observations required of a classification plot making the time required for data collection relatively modest in comparison.

Information about the observation of the plot that describes the event, such as the date, the persons involved, the geo-coordinates (including the datum and the precision of the record), the unique identifier of the observation, and the physical layout of the plot should be recorded as metadata. If the plot is observed more than once, it is important to separate data that are constant between measurements, such as geo-coordinates, from information particular to the observation event, such as date. A text description of the location is encouraged. The second group of observations contains facts about the site and its overall vegetation. Basic topographic information such as slope, aspect and elevation are nearly always collected. Most other environmental data are difficult to standardize, so these are usually tailored to the project or its larger context. For example, soil chemistry data can be very helpful for interpreting plots in a project, but results can vary greatly with protocol, and substantially even between labs using a consistent protocol, with the consequence that combining soils data from plots collected in separate projects must be done with great caution. Finally, summary records about the physical structure of the vegetation are often required, such as the height and cover of vegetation in different vertical strata. These seemingly simple measurements also vary significantly with protocol so care must be taken to retain consistency in data collection across a project and when integrating data from multiple projects.

Taxon identification and documentation present several challenges. Inevitably some taxa observed in the field will be unknown. As multiple taxa are often unknown, it is best to link a collection to a specific line number on the field data page so that future ambiguities are minimized. The taxon list should have each taxon recorded to the highest resolution possible, be it variety or family. Recognition of infraspecific types can prove invaluable during future data integration as varieties often migrate to full species status, and future splitting would not be possible without special information being recorded, such as variety name. Care should be taken to follow standard authorities for the taxa recognized and to record that authority (as opposed to the authority for creation of the name) so that in the future the meaning of the name can be evaluated. This step is necessary because the meaning of a taxonomic name can vary among treatments, and a taxon can have different names in different treatments owing to multiple, contrasting circumscriptions (see Franz *et al.* 2008, Jansen & Dengler 2010).

Each species in a plot is typically assigned a cover class value, and in many cases a cover class value is assigned specific to each stratum in which it occurs. Cover is the percent of the earth surface covered by a vertical projection of the leaves, though typically small holes within a single individual's crown are ignored. Cover class is an ordinal variable, typically with 5 to 10 possible values. Numerous scales have been proposed (summarized in van der Maarel 1979, Dengler *et al.* 2008 and Jennings *et al.* 2009, and see Table 1). The most frequently employed cover index is the 1-5 scale of Braun-Blanquet (1928) or some simple variant of it. Almost as common are variants of the 1-10 Domin scale (1928), such as that of Krajina (1933), the UK National Vegetation Classification (Rodwell 2006), the New Zealand Survey (Allen 1992) and the

Carolina Vegetation Survey (Peet *et al.* 1998). In selecting a cover scale, there are three important guidelines. First, it should be approximately logarithmic until at least 50% cover. This is because the human mind perceives cover in roughly a logarithmic way; we can perceive the difference between 1 and 2 percent, but not 51 and 52, as the first pair represents a doubling while the second is a small relative increase. Second, the index should be replicable between observers to the level that almost always two observers will be within one value of each other. Third, it is highly desirable that the scale be directly mapable onto the numeric units of the Braun-Blanquet scale to assure that datasets from diverse times and places can be integrated for at least some purposes.

4.5 Data preparation and integration

Once plot data have been collected, either in the field or from plot archives, it is necessary to integrate and standardize the data for analysis. This requires that inconsistencies in plot method, size, and taxonomy be addressed in a consistent and well-reasoned manner, with each step recorded for archiving with the database at the completion of the project.

4.5.1 Taxonomic integration

Construction of a taxonomically homogenous dataset can be a challenging task and typically requires investigator judgments on numerous inconsistencies. Because taxonomic adjustments will differ in their implications for data analysis, researchers should typically develop two datasets, one designed to address questions of species richness (species richness dataset) and one designed to address questions where between-plot similarity must be assessed (analysis dataset). In the species richness dataset, all entities recorded as different species in a plot should be retained as distinct, regardless of the taxonomic resolution. In the analysis dataset there should be a standard set of taxa used across all plots, and where taxa are inconsistently resolved they should usually be lumped together. If a small percentage of occurrences are reported only to the genus level, these taxon occurrences should be discarded from the analysis dataset; if most occurrences are unknown one should lump them to the genus. Taxa not resolved to at least the genus level should be dropped from the analysis dataset as such groups usually have little commonality in traits or distribution. If many taxa in a plot are not known to the species level, the plot should be dropped from the dataset. The trickiest cases are where many observations are known to species and still a significant number are known only to genus. What if in a dataset 70% of *Carex* occurrences are known to species and they span 20 taxa? Perhaps the numerous occurrences of *Carex* sp*.* should be dropped because they contain much less information than those identified to species, but the price is that there are missing records of shared taxa. Moreover, if one data source were consistently of lower resolution, there would be a signal attributable to a specific study.

Integrating taxon occurrences across datasets of mixed provenance presents greater uncertainty as to synonymy than does a single survey. Even when two occurrences are

unambiguously assigned the same taxonomic name, it is still necessary to verify that the taxonomic concepts are equivalent. This is because one taxonomic name can have many meanings in terms of specific sets of specimens, and a certain set of specimens could have many different names (Berendsohn *et al.* 2003, Jansen & Dengler 2011). Franz *et al.* (2008) describe the situation with the grass known as *Andropogon virginicus* in the Flora of the Carolinas (Radford *et al.* 1968), which when examined across eight taxonomic treatments reveals nine distinct sets of specimens variously arranged into 17 taxonomic concepts (combinations of the nine sets of specimens) and labeled with 27 scientific names. Thus, when plots from multiple sources report the presence of *Andropogon virginicus*, it is impossible to know how to combine them without knowledge of the taxonomic treatments the original authors followed, and even then there could easily be ambiguities requiring lumping to obtain unambiguous bins of taxon occurrences. The current situation in Europe serves to illustrate the mind-numbing complexity of integrating accurately and precisely across datasets. Schaminée *et al.* (2009) state that in order to establish the TurboVeg-based joint European vegetation database SynBioSys Europe (Schaminée *et al.* 2007), 30 national species lists with 300,000 names had to be mapped against each other. This mapping is strictly one of synonymy and different applications of names are in many cases not accounted for, leaving many potential traps for the unwary data aggregator.

4.5.2 Plot data integration across datasets

Compared to taxonomic integration, merging other aspects of plot records is relatively straightforward, even if somewhat arbitrary. For the most part there are only three major impediments: inconsistencies in cover scales, plot size and definition of vertical strata.

To the extent that cover scales nest into a small number of bins, such as those of the Braun-Blanquet scale, it is easiest to simply condense the number of bins. Where this nesting approach is not possible, one can convert the cover scale value to an absolute cover value and then back to a new cover scale value. In doing so the reader is advised to convert to the geometric mean of the range rather than the arithmetic mean as species occurrences tend to occur disproportionately in the lower portion of each cover class. Where no such conversions seem reasonable, analyses should be conducted with simple presence-absence data. In fact, some authors have argued that there is more interpretable information in presence-absence data than cover data because the degree of absence of a species cannot be known or readily estimated (Lambert & Dale 1964, Smartt *et al.* 1976, Wilson *submitted*, but see Beals 1984, McCune 1994).

Combining plots of different sizes in a single database is at best problematic. The variable most sensitive to plot size is species richness, but a rough correction can be achieved by adjusting the species richness of plots that are at most within a doubling of the target size by use of the species-area relationship (see Fridley *et al.* 2005). More problematic and uncertain are the implications of plot size differences for calculation of similarity and designation of species constancy and indicator species. The reduction in species number with decreasing plot size of necessity decreases similarity to larger plots with more species, and constancy is similarly

sensitive to plot size (Dengler *et al.* 2009). As a rule of thumb, all comparisons of richness should be made with plots of identical size, and all studies based on species similarity or constancy should be based on plots that do not range more than perhaps four-fold in area.

Most plot protocols call for recognition of vertical strata within a community, for which separate cover values are assigned for species. Unfortunately, these classes are not consistent between protocols. For example, the height cutoffs for strata can vary, and the actual definition of a stratum can vary from being based on the height of the individual plants (*e.g.* Mucina 2000) to simple vertical bands of leaf area (*e.g.* Allen 1992). Vertical strata can be combined for purposes of data integration and Jennings *et al.* (2009) suggest that a simple probabilistic calculation of species total cover across strata (C_i) can be calculated as

$$C_i = \left[1 - \prod_{j=1}^{n}\left(1 - \frac{\%cov\ j}{100}\right)\right] \times 100$$

assuming the leaf area in each stratum (%*cov j*) is statistically independent of the other strata.

4.5.3 Sampling intensity

The distribution of plot frequencies in phytosociological databases is far from even. Some types of vegetation have hundreds or even thousands of plots, whereas others may be represented by only a small number. In some forms of analysis the plots from the abundant types would dominate. Consequently, if we want an analysis to span the range of vegetation variation in the database, it may prove necessary to sample from the database in a stratified random fashion. Knollová *et al.* (2005) proposed several methods for stratifying phytosociological databases related to distribution along environmental or geographical axes, or relative to between-plot variation in species composition. Subsequently, a resampling method based on between-plot dissimilarity in species composition was proposed by De Cáceres *et al.* (2008). Lengyel *et al.* (2011) proposed a resampling method based on species composition where subsets of the database are selected randomly and the subsets with the lowest mean dissimilarity and lowest variance in similarity are retained for purposes of stratification.

4.6 Community entitation

In section 4.3 we distinguished between classification as the creation of classes *versus* the assignment of objects to classes. In this section we address the creation of classes or types from an undifferentiated data set of vegetation plots or relevés, which we will refer to as entitation — the creating of entities. Assignment of new vegetation plots to existing classifications is discussed as determination in section 4.9 below.

Vegetation scientists may have a broad range of ultimate objectives for classifying

vegetation (see section 4.1.1 above). From an operational perspective, however, the objective of vegetation classification is fairly simple — to create a set of vegetation types or syntaxa where (1) the types are mutually exclusive (no vegetation plot belongs to more than one type) and (2) the types are exhaustive (all vegetation plots are assigned to a type). Mathematicians call such a set of classes a "partition": every object is a member of strictly one set, and every set has at least one member. Perhaps not surprisingly, there is an extremely large number of ways to produce such a partition. In general, methods of vegetation classification can be characterized as expert-based *versus* algorithmic, with the algorithmic methods divided into numerical *versus* combinatorial.

4.6.1 Classification by table sorting

Vegetation classification by sorting of phytosociological tables has a long history in vegetation ecology, with methodological monographs from Braun-Blanquet (1928), Ellenberg (1956) and Becking (1957), with more recent treatments by Westhoff & van der Maarel (1973) and Mueller-Dombois & Ellenberg (1974, Chap. 9).

In table sorting methods the data on species occurrence or abundance by plot are organized in a rectangular matrix with species as rows and plots as columns. The objective is to order the rows and columns of the tables to create a block-structured table where abundances for individual species are concentrated in adjacent columns of a row, and species with similar distributions are concentrated in adjacent rows so that plots of similar composition occur in proximity in the table. Based on successive re-ordering of the rows and columns, the table can be divided into sections or blocks of co-occurring species with the blocks arranged in a diagonal down the table. Vegetation plots that include one (or more) of these blocks are assigned to the same syntaxon, and species that compose a given block are considered diagnostic of the syntaxon in which they occur. The specific meaning of diagnostic is the subject of considerable scientific development. Szafer & Pawlowski (1927), Becking (1957), Whittaker (1962), Westhoff & van der Maarel (1973) and Mueller-Dombois & Ellenberg (1974) distinguish "character species" based on fidelity of occurrence within classes and "differentiating (or differential) species" that are diagnostic in differentiating one class from another class while not necessarily being restricted to the focal class.

Table sorting by inspection was superseded many years ago by computer-aided approaches. The direct optimization of structured tables by iterative algorithms is difficult due to the extremely large number of possible solutions. The number of distinct table orderings is $n! \times m!$ where n is the number of plots and m is the number of species; even a simple table of 10 plots and 20 species has $10! \times 20! > 8.8 \times 10^{24}$possible orderings. Developing efficient numerical approaches to producing sorted tables thus became an area of active research (Westhoff & van der Maarel 1973).

In the last decade the development of computer-based or computer-aided table sorting has received renewed attention motivated in part by the need to manage tables of truly enormous size, such as when combining multiple national classifications into European-wide

classifications (Bruelheide & Chytrý 2000). Given the difficulty of direct optimization of large tables, most approaches have centered on statistical characterization of diagnostic species (see section 4.8.2 below). Among the more notable advancements was the development of the COCKTAIL algorithm for defining species groups by Bruelheide (2000). COCKTAIL starts with a preselected group of relevés or species and employs an iterative membership algorithm to refine the list of member species in each species group. Once no further candidate species are identified for membership in the type, a new type is begun from an alternative initial relevé or species group. Species fidelity to type is based on the *u* statistic (see section 4.8.1 below).

4.6.2 Numerical classification.

The most common approach to vegetation classification is by numerical means. Typically this requires defining a similarity or dissimilarity matrix among all the vegetation plots, and then clustering the plots into types. In operation, it is a three-step process of (1) defining (dis)similarity, (2) choosing a clustering algorithm, and (3) choosing the number of clusters revealed or desired. All three decisions strongly affect the results and have to be made in concert.

Dissimilarity and distance. There is an extraordinary number of dissimilarity/distance indices proposed or employed in vegetation ecology. Goodall (1973), Orlóci (1978), Hubálek (1982) and Legendre & Legendre (1998) all present comprehensive lists of indices that have been employed in community ecology; Mueller-Dombois & Ellenberg (1974), Kent (2012) and Ludwig & Reynolds (1988) emphasize shorter lists of commonly used indices. Confusingly, many indices have been independently derived and given more than one name. Other indices have a different name for the similarity index and its complement, the dissimilarity index, but vegetation ecologists often ignore the distinction and use the same name for both. Important distinctions among the indices concern the distinction between dissimilarity and distance and the use of presence/absence *versus* abundance data.

Dissimilarity and distance are similar concepts that characterize on a quantitative scale how different vegetation sample plots are from each other, but the mathematical bases of dissimilarity and distance are different. Dissimilarity is a set theoretic concept and represents the ratio of the disjunct elements of two sets (belonging to one or the other but not both) to the union of the two sets. Plots that have no species in common have a dissimilarity of 1, and plots that are identical have a dissimilarity of 0, with all other possibilities scaled [0,1]. Distance is a geometric concept and represents the sum of all the pairwise differences in abundance for species which occur in one or both plots. Identical plots have a distance of 0. Plots with no species in common have a distance determined by species richness (for presence absence indices) or standing crop (for quantitative indices) of the two plots; there is no upper bound. In practice a matrix is constructed with *n* rows and *n* columns for *n* vegetation plots where each row or column in the matrix expresses the dissimilarity or distance of one vegetation plot to all the other plots. Both dissimilarity and distance follow a set of axioms.

1) $d_{ii} = 0$, reflexive property; the dissimilarity or distance from a plot to itself is zero;
2) $d_{ij} = d_{ji}$, symmetric property; dissimilarity is independent of direction;

These two axioms are generally true of all dissimilarities or distances employed in vegetation ecology. Some, but not all, indices meet a third axiom.

3) $d_{ik} \leq d_{ij} + d_{jk}$, *i.e.* the dissimilarity or distance of a plot to another plot is less than or equal to the sum of the distances involving any third plot.

The third axiom is called the triangle inequality property and does not hold for many dissimilarity indices. Indices that meet all three axioms are "metric," and play a key role in analyses based in linear algebra.

Dissimilarity indices for presence/absence data often employ a 2x2 contingency table notation (Fig. 4.1). One of the earliest commonly used indices is the Steinhaus index (the complement of the Jaccard index): $(b + c)/(a + b + c)$; see Fig 4.1. This index can be viewed as the ratio of the number of species in one but not both plots to the pooled species list of the two plots. A commonly used alternative is the Marczewski index (the complement of the Sørensen index): $(b + c)/(2a + b + c)$, the ratio of the species in one but not both plots to the average number of species in the two plots. Both indices ignore *d*, the number of species in the data set that don't occur in either plot. Goodall (1973) and Legendre & Legendre (1998) argue strongly that ecologists should only use presence/absence dissimilarity indices that ignore joint absence (*d*).

For quantitative dissimilarity/distance indices, the abundance scale used can have a profound effect on the results. In vegetation ecology the scale is often not purely numeric (*e.g.* the Braun-Blanquet cover/abundance scale), and a lively debate has developed concerning the appropriate use of such data in quantitative analyses (Podani 2005, van der Maarel 2007). Nonetheless, most vegetation ecologists have adopted a pragmatic approach, and transform such scales to a numeric scale (see van der Maarel 1979, Noest *et al.* 1989). For scales with discrete classes of abundance, the widths of the intervals and the values chosen to represent each interval (often the interval midpoint but preferably the geometric mean of the endpoints) strongly affect the results. In general, linear abundance scales should be transformed to a convex, *e.g.* square root or log, scale that emphasizes differences for low values in the scale.

In addition to transformation, standardization of data can have a strong effect on results. Three standardizations are in common use in vegetation ecology: species maximum standardization, sample total standardization, and Wisconsin double standardization. Species maximum standardization divides the abundance of each species in each plot by the maximum value observed for that species in all plots in the data set, giving all species an equal voice in the calculation of dissimilarity/distance, and thus strongly de-emphasizing differences in dominance among sample plots. This can be useful where indicator species (section 4.8.2) exhibit low abundances and need increased weighting relevant to the dominants. However, this transformation can also increase the noise associated with rare species in the data and may work best where rare species are removed or down-weighted. In a comprehensive analysis of

the performance of different dissimilarity indices on simulated data, Faith *et al.* (1987) found that a species maximum standardization improved the performance of most indices; however, their simulated data may have contained disproportionately few rare species.

Sample total standardization divides the abundance of each species in a plot by the sum of abundances for all species in that plot. This transformation treats total abundance for each plot as equal, eliminating differences in productivity or standing crop among samples. This standardization can be effective when data were collected in different years or seasons, by different parties, or measured on different scales. Sample total standardization plays an important role when using geometric distances, such as Euclidean or Manhattan distance (see below). Geometric distances quantify the differences between plots without accounting for what the plots may have in common and can give a distorted perspective. A sample total standardization scales the differences relative to the total abundance and eliminates such problems. Some dissimilarity/distance indices (*e.g.* chord distance described below) have an inherent sample total standardization.

In Wisconsin double standardization, named for its use by Bray & Curtis (1957), data are first standardized by species maximum standardization and then by sample total standardization. Bray & Curtis's rationale for this sequence was that different life forms (trees *versus* non-trees) were measured on different scales and the species maximum standardization achieved a common scale. The subsequent sample total standardization corrected for the fact that not all plots had the same number of measurements.

Commonly used quantitative dissimilarity indices in vegetation ecology include the Bray-Curtis index (Table 4.2). The Bray-Curtis index has been criticized for not being a true metric (Orlóci 1978). However, in comparative tests it has often performed extremely well (Faith *et al.* 1987). Alternatively, the Marczewski-Steinhaus index (Table 4.2) is similar to the Bray-Curtis but is a true metric.

Geometric distances employed in vegetation ecology include Euclidean (Table 4.2 and Manhattan (or city-block) distance (Table 4.2). Euclidean distance is the common distance we employ to measure the distance between objects in our three-dimensional world and appears quite intuitive. Due to its use of squared abundances, however, it is quite sensitive to the range of abundances in the data. Manhattan distance is named for its similarity to walking distances in a city where all distances occur along the principal axes and travel along the diagonals is not possible. Both Euclidean and Manhattan distance benefit from a sample total standardization. Legendre & Gallagher (2001) examined the behavior of a number of dissimilarities and distances on artificial data and observed that Hellinger distance (Table 4.2) performed well at recovering ecological gradients. Hellinger distance can be viewed as the Euclidean distance of square root transformed sample total standardized data. Orlóci (1967, 1978) has demonstrated good results with chord distance (Table 4.2). Both Hellinger and chord distance employ inherent sample total standardization.

Hierarchical agglomerative clustering. Hierarchical agglomerative clustering algorithms begin with each vegetation sample in its own “cluster” and then iteratively fuse the least dissimilar clusters at each step. Ultimately, after *n-1* fusions (for *n* vegetation plots) all the plots are in a single cluster. The algorithms differ in how they define "least dissimilar" for clusters with more than one member (Table 4.3). Over the years many algorithms have been proposed and tested based in multidimensional geometry, graph theory, and information theory. We restrict our discussion to algorithms commonly employed in vegetation ecology.

In single linkage (nearest neighbor) clustering, the dissimilarity of two clusters is the dissimilarity between the two least dissimilar members of the respective clusters (Fig. 4.2a). This is similar to measuring the distance between islands or continents as the distance between the two closest points. As clusters get larger, there are more members you could be least dissimilar to, and existing clusters have a tendency to grow at the expense of starting new clusters. This leads to the phenomenon of “chaining” (Williams, Lambert & Lance 1966) where new vegetation plots are continually added to one large existing cluster. Due to the tendency to exhibit strong chaining, single linkage clustering is now rarely employed (Legendre & Legendre 1998, Podani 2000, McCune & Grace 2002).

In complete linkage clustering the dissimilarity of two clusters is the dissimilarity between the two most dissimilar vegetation plots of the respective clusters (Fig. 4.2b). This approach emphasizes maximum rather than minimum dissimilarity among clusters. As clusters get larger, there are more members to be potentially maximally dissimilar to, and joining existing clusters gets harder. This leads to more numerous, equally-sized often spherical clusters. In both the single linkage and complete linkage algorithms, the dissimilarity between clusters is decided by a single dissimilarity and the algorithms operate at the plot-level rather than the cluster-level (Williams, Lambert & Lance 1966). The algorithms are, therefore, sensitive to unusual plots or outliers.

In the average linkage method (also called UPGMA or Unweighted Paired Group using Averages, Sokal & Sneath 1963) method, the dissimilarity is the average dissimilarity of each plot in each cluster to all the plots on the other cluster (Fig. 4.2c). Average linkage performs intermediate to single linkage and complete linkage, *i.e.* it is less prone to chaining than single linkage, but may form irregularly-shaped clusters of varying size.

Ward's algorithm attempts to minimize the sums of squared distances from each plot to the centroid of its cluster, equivalent to variance minimization (Legendre & Legendre 1998). Beginning with every plot in its own cluster it fuses those clusters that result in the minimum increase in the sum of squared distances. Because it is based on a sum-of-squares criterion, the algorithm is most appropriately applied to a Euclidean distance matrix of plot dissimilarities (Legendre & Legendre 1998). However, many vegetation ecologists have been successful in applying Ward's algorithm to other dissimilarities such as Sørensen's (*e.g.* Wesche & von Wehrden 2011). Ward's algorithm tends to create compact spherical clusters where much of the variability in the dendrogram is compressed in the smaller clusters. This makes choosing relatively few large clusters rather easy, but sometimes hides considerable variability among the

more numerous smaller clusters.

Lance & Williams (1967) realized that many of the existing hierarchical agglomerative algorithms could be generalized to a single algorithm with specific coefficients in the among-cluster distance equation. This algorithm is mostly known today as “flexible-β” after one of the coefficients in the algorithm. Fig. 4.3d shows a flexible-β dendrogram with β set at the commonly employed value of -0.25. With this value (and suitable constraints on the other coefficients) flexible-β is intermediate to average linkage and complete linkage, and is generally recognized as a good compromise. By assigning β increasingly negative values (*e.g.* -0.5) the flexible-β algorithm more nearly approximates Ward's algorithm and provides an alternative that alleviates the concerns over requiring Euclidean distance.

All the hierarchical agglomerative algorithms initially produce a dendrogram that portrays the sequence of fusions into clusters of the sample plots. Dendrograms are aggregated from the bottom up. Early fusions of clusters in the algorithm constrain later fusions, and in hierarchical clustering the assignment of plots to clusters is never re-evaluated. Consequently, the relatively few clusters produced near the top of the dendrogram may show considerable artifact in plot assignment. While highly informative, dendrograms can be visually misleading as plots that are adjacent to each other but attached to different “branches” higher up may be quite dissimilar. An example is shown in Fig. 4.3 where the complete linkage and flexible-β algorithms produce what appear to be quite different dendrograms; re-ordering the plots along the horizontal axis would show that the solutions are very similar and the four cluster solutions (see below) are identical.

Dendrograms must be “sliced” to generate clusters of plots on the same “branch,” and the question of where to slice is a critical issue. Given an *a priori* desired number of clusters you can solve for the height at which to slice. Fig. 4.3 shows all four dendrograms sliced to produce four clusters. Often, however, the correct or desired number of clusters is not known, and we are interested in finding natural breaks in the dendrogram where the results are relatively insensitive to the precise height at which we slice. In the example shown natural breaks result in two or four clusters for complete linkage (Fig. 4.3b), two, three or five clusters for average linkage clustering (Fig. 4.3c) and two or three clusters for flexible-β (Fig. 4.3d). Further down in the dendrogram it is much more difficult to visually identify natural breaks and algorithmic approaches may be required.

Hierarchical divisive clustering. Hierarchical divisive clustering begins will all plots in a single cluster, which is then divided into two subclusters recursively until the clusters get too small or too homogeneous to subdivide according to criteria established by the user. Hierarchical divisive clustering algorithms are combinatorial, as opposed to numerical, and computing optimal results may be impossible. Accordingly, most divisive algorithms do not examine all possible solutions.

Two divisive algorithms are currently used in vegetation ecology: Two Way Indicator Species Analysis (TWINSPAN) and Divisive Analysis Clustering (DIANA). TWINSPAN (Hill 1979)

iteratively partitions the first axis of a Correspondence Analysis ordination (see Chapter 2). In practice the algorithm makes a number of *ad hoc* adjustments in choosing the exact point at which to partition at each step. Because TWINSPAN bifurcates each branch, the original algorithm always produces classifications where the number of classes is a power of two. Roleček *et al.* (2009) recently proposed a modification of the algorithm that employs a measure of cluster heterogeneity to determine which branches to split further. The result is a more natural classification with similar levels of cluster heterogeneity.

Kaufman & Rousseeuw (1990) introduced a hierarchical divisive algorithm, DIANA, that operates on a dissimilarity matrix. At each iteration DIANA identifies the cluster with the largest diameter (maximum within-cluster dissimilarity, equivalent to the complete linkage criterion). Within that cluster DIANA identifies the plot with the greatest average dissimilarity to all other plots in that cluster, and sets that plot aside as the seed for a "splinter group." All plots in the cluster that are more similar to the splinter group than the original cluster are then assigned to the splinter group, which forms a new cluster. Because DIANA is numerical, rather than combinatorial, it is fairly rapid but somewhat sensitive to outliers. Like hierarchical agglomerative algorithms, DIANA produces a dendrogram rather than clusters, and must be sliced to generate clusters. Because of the maximum diameter criterion, DIANA produces results most similar to complete linkage hierarchical agglomerative clustering.

Non-hierarchical partitioning algorithms. Non-hierarchical partitioning algorithms attempt to derive a specified number of clusters from an undifferentiated set of vegetation plots directly without a hierarchical dendrogram. In contrast to hierarchical approaches the number of clusters must be specified in advance. Non-hierarchical partitioning of objects into types is mathematically difficult due to the extraordinary number of possible solutions. For example, to classify only ten vegetation plots into non-overlapping types there are 118515 possible solutions. To simplify finding good solutions to this problem many non-hierarchical algorithms search for suitable "seeds" to start each cluster and then assign each vegetation plot to the nearest seed. The original approach was called the k-means algorithm (Hartigan & Wong 1979), which minimized the sum of squared distances between points and the centroid of the cluster to which they were assigned. Modifications of the algorithm generally involve methods to choose the initial seeds and iteratively re-designate seeds. The k-means algorithm is strongly biased to create circular clusters of equal size rather than identifying natural discontinuities in the data. In addition, the algorithm is sensitive to the initial choice of seeds, and often requires multiple independent starts to ensure a good (although not necessarily optimal) solution.

Kaufman & Rousseeuw (1990) introduced a variation on k-means clustering called Partitioning Around Medoids (PAM). In the PAM algorithm, the seed for cluster formation (the medoid) represents an actual plot, called the representative object, rather than a geometric centroid. A deterministic algorithm selects the initial medoids, and because PAM does not require calculating centroids, it can operate on a broad range of dissimilarity indices other than Euclidean distance. Roberts (2010) defined two iterative non-hierarchical partitioning algorithms called OPTPART and OPTSIL. OPTPART iteratively reassigns plots to clusters to maximize the ratio of within-cluster similarity to among-cluster similarity. OPTSIL iteratively

reassigns plots to clusters to maximize the similarity of a plot to its assigned cluster compared to the next most similar cluster (see section 4.7.2 below for more detail). Fuzzy clustering algorithms have also been proposed as an alternative to non-hierarchical algorithms wherein plots can have partial membership in multiple types (Equiha 1990, Podani 1990, De Cáceres *et al.* 2010a). These approaches recognize that not all plots are representative of a single type and sometimes are intermediate to clearly recognized types, but the resulting classification structure is more complex.

Non-hierarchical partitioning methods suffer from the requirement that the number of clusters to be solved for must be specified in advance. They can also be slow to converge to a solution for some data sets. In practice, it is generally necessary to try multiple starts for a variety of cluster numbers and to compare the results to identify the best solution based on cluster validity statistics (section 4.7), cluster characterization based on ancillary data (section 4.8), or synthesis tables of the clusters. On the other hand, non-hierarchical partitioning algorithms generally do not suffer from the artifact of fusion sequences constraining results as all plots are re-examined for best fit at each iteration.

4.7 Cluster assessment

The two objectives of assessing vegetation classes derived from any clustering method are to assure that (1) types are relatively homogeneous and distinct from other types, and (2) distributions of species within types exhibit high fidelity and ecologically interpretable patterns. Assessing the goodness of clustering ("cluster validity") is a vast field with a voluminous literature. Aho *et al.* (2008) present a recent review of cluster assessment methods for vegetation classifications. These authors distinguish geometric evaluators based on dissimilarity matrices *versus* non-geometric evaluators based on species distributions within clusters, often with a view to identifying diagnostic species. Some methods attempt to measure structure in a vegetation table directly.

4.7.1 Table-based methods

Feoli & Orlóci (1979) proposed a method termed Analysis of Concentration (AOC) to assess the structure of vegetation tables based on the density of non-zero values within species and sample blocks recognized by the vegetation ecologist. Blocks with high density (dominated by the presence of species in plots within the block) and blocks of low density (dominated by the absence of species in plots within the block) are compared to a random expectation by χ^2 analysis. Deviation from expectation is a direct measure of the degree of structuring of the table, and it is possible to scale the divergence to a relative scale of [0,1]. Many of the optimization criteria from iterative table-sorting algorithms (*e.g.* Podani & Feoli 1991, Bruelheide & Flintrop 1994) can be used to measure the quality of the final results even when that algorithm was not employed to define the classes. Generally these statistics are insensitive to the ordering of species or plots within blocks, but still measure cluster structure from the table.

4.7.2 Dissimilarity-based methods

Dissimilarity-based methods of cluster assessment operate on dissimilarity matrices, and can be applied whether numerical clustering was employed in defining the types or not. Aho *et al.* (2008) refer to these approaches as geometric evaluators and list five statistics useful in assessing goodness of clustering: Average Silhouette Width (Rousseeuw 1987), C-Index (Hubert & Levin 1976), Gamma (Goodman & Kruskal 1954), the PARTANA ratio (Roberts 2010, Aho *et al.* 2008), and Point Biserial Correlation (Brogden 1949). Two of these indices are highlighted below.

Rousseeuw (1987) defined silhouette width as a measure of the degree to which plots are more similar (less dissimilar) to the type to which they are assigned than to the most similar alternative type. Positive values indicate a good fit, and negative values indicate samples more similar to another cluster than to the cluster to which they are assigned. Thus, the quality of each cluster can be assessed by the mean silhouette widths of all plots assigned to that cluster and the number of negative silhouette widths, and the overall quality of the classification can be assessed by the global mean silhouette width and the number of negative silhouette widths. Silhouette width is a "local" evaluator in the sense that each plot is only compared to the single other cluster to which it is least dissimilar regardless of the number of clusters. That comparative cluster may be different for every plot within a cluster. Roberts (Roberts 2010, Aho *et al.* 2008) defined a dissimilarity-based statistic called the PARTANA (PARtition ANAlysis) ratio that calculates the ratio of the mean similarity of plots within types to the mean similarity of plots among types. Good clusters have a high within-cluster similarity and low among-cluster similarities, and plots that fit well within their cluster have a higher mean similarity to their cluster than to other clusters. In contrast to silhouette width, PARTANA is a global statistic that compares every cluster to every other cluster.

4.7.3 Indicator species methods.

Statistical analysis of diagnostic or indicator species is often used as an evaluator of clustering effectiveness. The IndVal statistic (Dufrêne & Legendre 1997, and as modified by Podani & Csányi 2010, see below) and the OptimClass approach of Tichý *et al.* (2010) have both been effectively used in selecting "optimal" solutions from competing alternative classifications. However, as the identification of diagnostic and indicator species is of significant interest in community characterization, it is treated in section 4.8.

4.8 Community characterization

Once a set of types has been developed, it is desirable to develop concise representations of the compositional and ecological characteristics of the types. The data often represent a large number of species and plots, as well as possible environmental attributes, and efficient

summaries are required for effective communication.

4.8.1 Synoptic tables

One common and simple approach is to produce a synoptic table for the types recognized with species as rows, types as columns, and values of frequency, mean abundance, or preferably both, for each species in each type entered into the table. In U.S. vegetation classifications such tables are often called constancy/abundance tables. Similar to the more expansive structured tables described above (see section 4.7.1), the species (table rows) are often ordered to highlight the diagnostic species of the types. In large data sets with numerous types, even the synoptic tables can get quite large and unwieldy.

4.8.2 Diagnostic and indicator species

Deriving statistical indices of diagnostic or indicator species has been an area of significant activity in the last decade. Here we distinguish two groups of approaches: probabilistic *versus* composite. In general the probabilistic approaches calculate the "fidelity" of species to types or clusters based on presence/absence data and evaluate the deviation of species occurrence within types from a random distribution of taxa. Generally, each type or class is considered individually against all the other types pooled. The composite approach combines fidelity and the distribution of a species' abundance across types to create a single index. Because the null distribution of this index is not known, the deviation from expectation for the index values has to be estimated by permutation techniques.

Juhász-Nagy (1964) in De Cáceres *et al.* 2008 described three aspects of species fidelity that influence the indices in use today : (Type I) the occurrence of a species typically only within a vegetation type, although it may not occur in all (or even most) plots within the type; (Type II) the commonness or ubiquity of a species within a type although the species may be widespread outside the type; and (Type III) joint fidelity where a species occurs primarily within a single type and occurs in all (or most) plots within that type. The first case we might call "sufficient" in that the occurrence of that species is sufficient to indicate the type, the second case we might call "necessary" in that if you are in that type you should see that species, and the third case we might call necessary and sufficient.

4.8.3 Probabilistic indices of species fidelity

The general approach to probabilistic identification of diagnostic species is to calculate an index of concentration, and then the probability of obtaining as high or higher a concentration of a species within a given type as observed. For simplicity these indices are generally calculated on presence/absence data and concentration is calculated as number of occurrences (though see Willner *et al.* 2009). The most common approach is to produce a 2x2 contingency table of occurrences of a species in a type and calculate the ϕ index (Sokal & Rohlf 1995: 741, 743). Following notation established by Bruelheide (2000), the analysis is as follows:

N = total number of sites

N_p = number of sites in type of interest
n = number of occurrences of species of interest
n_p = number of occurrences of species in type

$$\Phi = \frac{N \times n_p - n \times N_p}{\sqrt{n \times N_p \times (N-n) \times (N - N_p)}}$$

Φ takes values in [−1, 1] reflecting perfect avoidance to perfect concordance of the species in the type. The statistical significance of the index can be calculated from Fisher's exact test. Bruelheide (2000) proposed that for species that occurred more than ten times a normal distribution approximation could be used, calculating

$$u = \frac{n_p - \mu}{\sqrt{n \times N_p/N \times (1 - N_p/N)}}$$

dividing the observed number of occurrences n_p minus the expected number of occurrences ($\mu = n \times N_p/N$) by the standard deviation of the binomial, preferably after applying a continuity correction to the numerator. Chytrý *et al.* (2002) preferred to divide by the standard deviation of a hypergeometric random variable and called the resulting value u_{hyp}.

$$u_{hyp} = \frac{n_p - \mu}{\sqrt{n \times N_p \times (N-n) \times (N - N_p)/(N^2 \times (N-1))}}$$

In either case the index of fidelity is scaled in units of standard deviation from expectation, rather than [-1,1]. As we are primarily interested in positive values of the index, a one-tailed test of significance can be performed on the index.

Chytrý *et al.* (2002) compared a range of statistical indices (including ϕ, u and u_{hyp}) for use in identifying diagnostic species on a classified data set from dry grasslands in Czech Republic. Rankings achieved by the probabilistic indices were very similar, although correction for continuity tended to reduce the values for rare species. Tichý &Chytrý (2006) argued that fidelity indices such as ϕ are sensitive to variability in the size of types or clusters, and proposed a modification of the ϕ coefficient that normalizes cluster size. The number of occurrences for a species and the number of occurrences within the type of interest are rescaled to a constant cluster size while maintaining the ratio of within-type to out-of-type occurrences. The new equalized ϕ values are comparable across clusters of different sizes. By adjusting the size of the normalized cluster relative to the total number of plots, the index can be made more or less sensitive to rare species relative to more common species. Normalized ϕ values are not appropriate for testing of statistical significance, so significance testing should occur before normalizing. Alternatively, the data can be sub-sampled to equal cluster sizes before the analysis (see section 4.5.4).

De Cáceres *et al.* (2008) present a detailed discussion of the importance of context in

identifying diagnostic and differential species. Willner *et al.* (2009) studied a range of fidelity indices on real data and found that differences in context were more important than the use of different indices of fidelity in identifying diagnostic species. The range of other vegetation types considered strongly influences the determination of species values. Approaches that compare the presence of species within types compared to outside the type can find character species with high fidelity, but miss many differential species that are not globally differential. A solution proposed long ago by Goodall (1953) consists of comparing the distribution of species within types to the type where the species is next most common.

Most of the numeric approaches to identifying indicator species focus on species with high fidelity as opposed to differential species. Tsiripidis *et al.* (2009) developed a method based on taxon relative constancy within types to identify differential species directly. While the algorithm is somewhat *ad hoc*, it proved successful in application to both simulated and actual data, and is logically related to thresholds employed in more classical phytosociology. Alternatively, a statistical numeric approach to identifying differential species is to use classification trees (Breiman *et al.* 1984) or random forest classifiers (Breiman 2001) on the plot-level compositional data to identify species useful in predicting the membership of plots in types (see section 4.9.1 below).

4.8.4 Composite Indices

The most widely employed statistic for identifying diagnostic species in a classification is Dufrêne & Legendre's (1997) IndVal statistic. Using the notation introduced by Bruelheide (2000, see above**)**

$$IV_{ip} = A_{ip} \times B_{ip} \times 100 \quad \text{where} \quad A_{ip} = \frac{\sum_{j \in p} a_{ij}/N_p}{\sum_c^C \sum_{j \in c} a_{ij}/N_c};\ B_{ip} = \frac{n_p}{N_p}$$

where IV_{ip} is the indicator value of species *i* to cluster *p*, a_{ij} is the abundance of species *i* in plot *j*, *c* is a cluster from one to *C* clusters, and N_c is the number of plots in cluster *c*.

The first term (A_{ip}) is the average abundance of the species in plots in the cluster of interest divided by the sum of the average abundances in all clusters. Calculating the sum of averages is an unusual calculation, but in this case it makes the relative abundances independent of cluster size. The second term (B_{ip}) is simply the relative frequency of the species in the cluster (Type II fidelity, as given above).

To achieve maximum indicator value a species must occur in every plot assigned to that type and no plots outside the type. Species that are restricted to a single type, but which occur in only a subset of the plots assigned to that type, are given an indicator value equal to their frequency; species that occur in every plot of the type, but which also occur in other types, are assigned an indicator value proportional to their relative average abundance within the type. The values are tested for statistical significance by permutation. The IndVal statistic attempts to find species that are both necessary and sufficient (*i.e.* if you see the species you should be in

the indicated type, and if you are in the indicated type the species should be present). As a comparative metric of overall classification efficacy, Dufrêne & Legendre proposed summing the statistically significant indicator values across species, or alternatively counting the number of significant indicator species and choosing the partition that maximizes the statistic.

The dual requirements that indicator species have high frequency in the indicated type and low abundance outside the type bias the IndVal statistic in favor of species that occur in the data at approximately mean cluster size. However, widespread species can have compact, ecologically informative distributions occurring with high fidelity in pooled types that are adjacent along gradients. De Cáceres *et al.* (2010b) developed a modified IndVal statistic that pools types into all possible larger groups and calculates the IndVal statistic (as well as the point biserial correlation) for those groups. Species with wider niche breadths could, thus, be recognized as indicative of a union of possibly several types.

Podani & Csányi (2010) noted that the first term of IndVal as given above is independent of the number of types being considered and represents concentration as opposed to specificity. They argued that specificity should consider how many types are in the data set and proposed a modification comparing the difference of the average abundance of a species in the type minus its average abundance in all other types, normalized by the maximum average abundance for the species in any type. This has the effect of changing the scale of indicator value from [0,1] to [-1,1], where species have negative specificity to types where their average abundance is less than their average abundance in all types. Ecologists have argued for years whether or not the lack of species can be diagnostic, but Podani & Csányi note their proposed index is consistent with the position of Juhász-Nagy (1964) that the absence of a ubiquitous species can be indicative.

4.9 Community determination

Determination is the assignment of a plot to an existing type based on comparison with the typical composition of candidate types. Determination may be absolute (or crisp) where the plot is assigned to only a single type, or fuzzy where the plot is given grades of membership in multiple types (De Cáceres *et al.* 2009, 2010a). The USNVC and VegBank allow five possible levels of determination: Absolutely Wrong, Understandable but Wrong, Reasonable or Acceptable Answer, Good Answer, and Absolutely Right (Gopal & Woodcock 1994). Alternatively, fuzzy set theory can be employed, where plots are assigned memberships in types in the range [0,1], typically where the sum of all memberships must equal one. Van Tongeren *et al.* (2008) rank the potential types in order of fit from 1 to 10 whereas De Cáceres *et al.* (2009) noted first and second best fit. In a manner similar to entitation, determination can be based on either actual compositional data or on (dis)similarities calculated among plots, or both.

Developing numerical or combinatorial approaches to assign plots or correct plot assignment is exceedingly difficult. For large data sets the number of plots and the number of

types is large and the dimensionality of the problem is typically very high. However, given the importance of developing comprehensive vegetation classifications, efforts to perfect such algorithms will certainly be given high priority by vegetation and computer scientists.

4.9.1 Expert-based approaches

Type membership for plots is often determined by expert opinion. Experienced vegetation ecologists employ an understanding of data context and intuitive species weighting in selecting the appropriate type for a plot. Often when numeric approaches are employed the results are validated using determinations by experts (treated as "truth"). However, as noted by van Tongeren (2008), mistaken determination by experts is a source of error unaccounted for in tests of numeric methods. Perhaps more importantly, as noted by Gégout & Coudin (2012), given the size of the task of producing national or regional classifications, there simply aren't enough experts to accomplish the task.

4.9.2 Dichotomous Keys

Dichotomous keys are extremely useful tools for field determination of new plots or relevés, as long as the list of possible types is not too long. Automated procedures for generating dichotomous keys are available using classification trees (Breiman *et al.* 1984) or random forest classifiers (Breiman 2001) on plot-level compositional data. However, given the stochastic nature of species distributions, dichotomous keys are limited by using the abundance of a single species (or a few pooled species) at each decision point, rather than a more synthetic perspective. In addition, dichotomous keys (and the differential species identified by them) are limited by context. If a type is widespread, then the differential species may vary by region, and application of a dichotomous key outside the region where the calibration plots were collected may prove highly error-prone. Keys must be recognized as useful but fallible tools for narrowing down the list of candidate types (Pfister *et al*. 1977, Rodwell 2006). Users must still compare the composition and environmental attributes of the indicated type and similar types to make a clear determination.

4.9.3 Numeric Approaches

Ĉerná & Chytrý (2005) employed the ϕ index (see section 4.8.3) in an application of neural nets (multilayer perceptron) to predict plot membership in 11 *a priori* alliances for 4186 relevés of Czech grasslands. The neural net was fit to a subset of the relevés (the training set), limited from over-fitting by another subset of relevés (the selection set) and tested on a third set of relevés (the test set). When the training data set was randomly selected from the pool of relevés, the neural net obtained from 80.1 to 83.0% correct assignment of the test data. Surprisingly, when the training data were selected by emphasizing relevés with high numbers of diagnostic species, the accuracy declined to 77.0 to 79.6%. Ĉerná & Chytrý regard the use of neural nets for plot assignment as promising, but note that the model is essentially a black box and does not produce keys useful for field application.

Gégout & Coudin (2011) also employed the ϕ index (see section 4.8.3) to develop a model for assigning plots to pre-existing types. ϕ was calculated for every species in every type using the data from the original (calibration) plots. Then the fidelity of a plot to a type (F_{ij}) was calculated as the mean ϕ for all species in the plot to that type.

$$F_{ij} = \sum_{k=1}^{n} \varphi_{kj}/n$$

This fidelity was compared to the mean fidelity of all plots used to define that type.

$$A_{ij} = \left(F_{ij} - \overline{F_j}\right)/s\left(F_j\right)$$

where A_{ij} = the affinity of plot I to type j, $\overline{F_j}$ is the mean fidelity for all plots in type j, and $s\left(F_j\right)$ is the standard deviation of F_j. Plots were assigned to the type for which they had the highest affinity. There was a 60% agreement of assignment to type compared to assignment by phytosociological experts on the calibration plots. For 800 plots independent of those used to define the types, agreement with expert assignment dropped to 47%.

Van Tongeren *et al*. (2008) developed a numerical determination approach called ASSOCIA based on a composite index combining presence/absence data and abundance data using weighted averaging. For the presence/absence data the deviance (-2 ln(likelihood)) associated with plot membership of a plot to a type is calculated for all possible types. For the abundance data a modified Euclidean distance is calculated from the plot to the centroid of all types. This approach has the significant advantage that it can employ synoptic tables, as opposed to full plot-level data, thus allowing comparisons to published classifications where the raw data are not available.

De Cáceres *et al.* (2009, 2010a) explored fuzzy approaches to determination. While the fuzzy classifiers performed well in general, they proved susceptible to poorly defined types in the set of possible choices, and differed in their response to outliers as opposed to intermediate plots.

4.10 Classification integration

With the growing importance of large, comprehensive classification systems such as that of the Braun-Blanquet system and the U.S. National Vegetation Classification, it is critical that new classification work be integrated into a broader framework. Additionally, existing classifications need to be reconciled into an integrated framework to achieve a consistent, comprehensive system (Bruelheide & Chytrý 2000, De Cáceres & Wiser 2012). There are significant challenges to achieving such an integrated system. There are four components to managing classifications (De Cáceres *et al.* 2010a): (1) assigning new relevé data into existing types, (2) updating the

types to reflect the additional data, (3) defining new types for plots that don't fit the current classification, and (4) reconciling and validating the modified classifications. De Cáceres & Wiser (2012) provide guidelines to ensure that the products of vegetation classification efforts can be integrated into broader classification frameworks, modified and extended in the future, and can be used to communicate information about vegetation stands beyond those included in the original analysis.

Here is a simple overview of the problem of integration. If all of the vegetation plots that are characteristic of a classification unit under one system (say A) would be assigned to a single classification unit of another system (say B), we will say that the relationship (or mapping) is one-to-one. If, however, plots that define a type in classification A would be assigned to more than one type in classification B, we will say the mapping is one-to-many. If the mapping is one-to-many in both directions the classifications are significantly different and reconciliation will be difficult, for the same reasons as mapping of taxonomic concepts for plants is challenging.

4.10.1 Classification resolution

Most vegetation classifications are hierarchical with lower levels nested into broader types. It makes sense to begin the discussion of classification integration at the lowest practical level, the association, as upper levels are often defined in terms of their component lower units (but see the USNVC for a combined bottom-up and top-down system). We refer to the heterogeneity of vegetation within an association (how finely divided into types the vegetation is) as classification resolution. If classifications to be reconciled differ significantly in resolution, then a one-to-one correspondence cannot be established. The best case is that in one direction the mapping is one-to-many and in the reverse direction it is one-to-one; in this case one-to-one mapping may still be achieved by lumping the more finely resolved types or splitting the more coarsely resolved types. Given a standard definition for intra-association heterogeneity, this would be a simple decision. However, no standard currently exists (although Mueller-Dombois & Ellenberg 1974 suggested that all plots within an association should have a Jaccard's similarity index of at least 25% to the typal plot). The variability in association resolution across classifications could be used to guide this decision.

4.10.2 Classification alignment, precedence and continuity

Even given similar levels of classification resolution between two adjoining classifications, it is likely that the classifications will still exhibit one-to-many relationships in both directions. Recurrent patterns of vegetation composition (associations) are determined in part by the pattern of landscapes acting on the regional species pool (Austin and Smith 1989). In an adjoining region differences in these landscape patterns may create different recurring community patterns from the same species pool. In these cases it may be necessary to pool the plot data from both areas and seek new associations that better represent the larger-scaled pattern of community composition and distribution. Similarly, a detailed study of a narrowly circumscribed geographic region (perhaps a national park) may yield an intuitively very satisfying classification that does not map well onto a geographically broad classification (say

one for all of Europe or the U.S.). In these cases is will be necessary to be cautious in proposing changes in the larger-scale classifications so as to avoid disharmonies in application of the classification in other regions.

Vegetation classifications represent significant scientific achievements often accomplished by a large number of people over a long period of time. Much of the utility of the classification, however, is tied to the information content of the classes. Often important ancillary information on productivity, animal habitat suitability, conservation priority and hazards are associated with each unit in the classification by accumulated experience or specific monitoring or research programs. Maps of classification unit distribution may feature prominently in land-management activities. Significant revisions of existing classifications run the risk of making such information obsolete. Accordingly, while new methods or new data or the desire to reconcile with adjacent areas sometimes lead to revised classifications, this should be done sparingly. At a minimum, considerable effort should be given to documenting the mapping from old types to new.

4.10.3 Cross-referencing classifications

An alternative approach to aggregating classifications into new systems is to develop a formal cross-referencing system that identifies synonymy among classifications as described above. One approach is a set theoretic system that follows the international standard for taxonomic mapping (TDWG 2005) in defining the relationship of each type in one classification with each type in another as: (1) is conguent, (2) is contained in (3) contains, (4) intersects with, or (5) is disjunct from, as is implemented for community classification in the VegBank archive (Peet *et al.* 2012). By knowing the relationship of a type in one classification to all types in another classification it is possible to erect higher order relationships by network algorithms. Such an approach preserves the ancillary information associated with types in legacy classifications and minimizes unnecessary dynamics in the larger classification enterprise. On the other hand, it imposes additional complexity on regional efforts.

4.10.4 Nomenclature

Each of the major vegetation classification systems has its own nomenclatural rules. The best established and most detailed is the International Code of Phytosociological Nomenclature, which applies to units in the Braun-Blanquet system (Weber *et al.* 2000) and is maintained by the International Association for Vegetation Science. This system is modeled after the nomenclature rules for plant taxa (Dengler *et al.* 2008). Among several names for a syntaxon, the oldest validly published name has priority, and each syntaxon name is connected to a nomenclatural type (a single plot for associations, or a validly described lower-rank syntaxon in the case of a higher syntaxa), which determines the usage of the name. Syntaxon names are based on the scientific names of one or two plant species or infraspecific taxa that usually are characteristic in the particular vegetation type. An ‘author citation’ (*i.e.* the author(s) and year of the first valid publication) also forms part of the complete syntaxon name.

The US National Vegetation Classification (http://usnvc.org) has less formal naming rules. Each association and alliance is assigned a scientific name based on the names of plant species that occur in the type (Jennings *et al.* 2009). Dominant and diagnostic taxa are used in naming a type and are derived from the tabular summaries of the type. The number of species names in the name can vary from one to five, with those predominantly in the same stratum separated by a hyphen (-), and those predominantly in different strata separated by a slash (/).Association or alliance names include the term Association or Alliance as part of the name to indicate the level of the type in the hierarchy, as well as a descriptive physiognomic term, such as forest or grassland.

4.11 Documentation

4.11.1 Publication

Publication is critical for disseminating the results of vegetation classification research, though it plays different roles in different classification systems. In the Braun-Blanquet system vegetation types are defined in publications, much as species are. Typically, these publications contain synoptic tables with species as rows and communities as columns. For classification publications constructed outside the framework of the Braun-Blanquet system, tabular summaries are still important, but less emphasis is placed on sorting or identification of diagnostic species. More typically, the most characteristic species are indicated. One effective manner of doing this is by including only the prevalent species, defined as the 'n' most frequent species, where 'n' is the average number of species per plot (Curtis 1959). In addition, it is common to flag the species with high indicator value as defined by some standard metric, such as that of Dufrêne & Legendre (1997).

4.11.2 Plot archives

With the advent of inexpensive digital archiving of data and widespread access to digital archives over the web, there is a growing expectation that key original data will be made available in permanent public archives (Jones *et al.* 2006, Vision 2010). As a consequence, analyses can now be redone with expanded datasets or with different methodologies, and new questions can be asked through use of large quantities of available data. This trend toward archiving original data is particularly important for vegetation classification initiatives. Large national and multi-national classifications need to evolve, and this is only possible if plots records are permanently archived, much like systematics depends on museum collections that have been examined and determined by a series of monographers. The USNVC now requires that plot data used to advance the classification be made available in public archives. Already in excess of 2.4 million vegetation plots are reported in the Global Index of Vegetation Databases (GIVD; Dengler *et al.* 2011), a significant proportion of which is publicly available.

Key to efficient reuse of data is that the records conform to some standardized format. The widespread use of TurboVeg (Hennekens & Schaminée 2000) as a database for plots consistent with the Braun-Blanquet approach has meant that millions of plots can be exchanged in an efficient manner. However, TurboVeg supports only a very limited range of plot types and formats. To solve this problem Veg-X has been proposed as an international data exchange standard for vegetation plots of nearly all formats (Wiser *et al.* 2011). Widespread application of the Veg-X format would greatly simplify both sharing of data and ease of application of software tools.

4.12 Future directions and challenges

Given the pressing need for documenting and monitoring the Earth's biodiversity and for providing context for broader ecological research, vegetation classification has received increasing attention in recent decades in both academic ecology and across a broad range of user communities. This new and broader set of applications also suggests that we need to move beyond individual and idiosyncratic classifications toward large, consensus classifications that combine the effort of many persons to produce and maintain a unified and comprehensive whole, subject to revision in an open and transparent manner. Toward this end individual workers should conform to established standards for collecting and archiving plot data. Not only will this significantly advance vegetation classification, but it will also facilitate future international collaboration and synthesis.

Computer databases and numerical approaches will become increasingly important for developing large consensus classifications. While a single preferred protocol is unlikely to emerge, increased testing of competing approaches on large regional or national classifications should provide insights into the task-specific utility of each approach. Transparent algorithms should be strongly preferred, although the specific nature of vegetation means that special-purpose software may still be required. As emphasized by De Cáceres &Wiser (2012), formal rules for assigning plot data to specific types will play an increasingly important role.

Vegetation scientists need access to the data employed in vegetation classifications. Numerous plot databases currently exist (Dengler *et al.* 2011), and progress is being made on data transfer protocols that will facilitate access to and utility of such data (Wiser *et al.* 2011). The development of better tools for managing and analyzing the massive vegetation data sets anticipated in future classification efforts is an area of active research and development.

The greatest future challenge may be integrating the numerous existing classifications into a comprehensive system. The USNVC has included a peer review protocol for modifying the classification from the founding of the system. Ironically, the US may benefit in this effort from the historical lack of emphasis on vegetation classification in North America, beginning from almost a clean slate. The long legacy of vegetation classification in Europe means that many more vegetation types are formally recognized. Thus, reconciliation of existing classifications will play a much larger role in Europe than in the US.

Vegetation is complex and dynamic and efforts to characterize it in a formal structure are inherently problematic. Nonetheless, identifying those problem areas focuses the efforts of vegetation science into new research areas of interest to a broad range of scientists in complexity science, database design, multivariate analysis, expert systems, and many other fields.

References

Aho, K., Roberts, D.W. & Weaver, T. (2008) Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. *Journal of Vegetation Science* **19**, 549-562.

Allen, R. B. (1992) RECCE: an inventory method for describing New Zealand's vegetation cover. *Forestry Research Institute Bulletin* 176. FRI, Christchurch, New Zealand.

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, USA.

Austin, M. P. (2012) Vegetation and environment: discontinuities and continuities. Chapter 2, this volume.

Austin, M.P. & Heyligers, P.C. (1989) Vegetation survey design for conservation: Gradsect sampling of forests in Northeastern New South Wales. *Biological Conservation* **50**, 13-32.

Austin, M.P. & Heyligers, P.C. (1992) New approaches to vegetation survey design: Gradsect sampling. In: Margules,C.R. & Austin, M.P. Eds., Nature conservation: cost effective biological surveys and data analysis. Chapter 5, CSIRO, Australia.

Austin, M.P. & Smith, T.M. (1989) A new model for the continuum concept. *Vegetatio* **83**, 35-47.

Bailey, R. G. (1976) *Ecoregions of the United States* (map). U.S. Forest Service, Intermountain Region, Ogden, Utah, USA.

Beals, E.W. (1984) Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research* **14**, 1-55.

Bechtold, W.A., & Patterson, P.L. (2004) The enhanced Forest Inventory and Analysis Program--national sampling design and estimation procedures. *General Technical Report* SRS-80. U.S. Department of Agriculture Forest Service, Southern Research Station, Asheville, North Carolina.

Becking, R.W. (1957) The Zurich-Montpellier school of phytosociology. *Botanical Reviews* **23**, 411-488.

Berendsohn, W.G., Döring, M., Geoffroy, M., Glück, K., Güntsch, A., Hahn, A., Kusber, W.-H., Li, J., Rüpert, D. & Specht, F. (2003) The Berlin Model: a concept-based taxonomic information model. In: Berendsohn, W.G. (ed.) *MoReTax – handling factual information linked to taxonomic concepts in biology* [*Schriftenreihe für Vegetationskunde*, no. 39]. pp. 15–42. Federal Agency for Nature Conservation, Bonn, DE.

Box, E. & Fujiwara, K. (2012) Vegetation types and their broad-scale distribution. Chapter 15, this volume.

Braun-Blanquet, J. (1928) *Pflanzensoziologie: Gründzuge der Vegetationskunde*. Springer-Verlag, Berlin, Germany.

Braun-Blanquet, J. (1964) *Pflanzensoziologie*. 3rd edition, Springer. 364 pp.

Bray, J.R. & Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27**, 326-349.

Breiman, L. (2001) Random forests. *Machine Learning* **45**, 5-32.

Breiman, L., Friedman, J., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth.

Brogden, H.E. (1949) A new coefficient: application to biserial correlation and to estimation of selective efficiency. *Psychometrica* **14**, 169-182.

Bruelheide, H. (2000) A new measure of fidelity and its application to defining species groups. *Journal of Vegetation Science* **11**, 167--178.

Bruelheide, H. & Chytrý, M. (2000) Towards unification of national vegetation classifications: A comparison of two methods for analysis of large data sets. *Journal of Vegetation Science* **11**, 295-306.

Bruelheide, H. & Flintrop, T. (1994) Arranging phytosociological tables by species-relevé groups. *Journal of Vegetation Science* **5**, 311-316.

Černá, L. & Chytrý, M. (2005) Supervised classification of plant communities with artificial neural networks. *Journal of Vegetation Science* **16**, 407-414.

Chytrý M. & Otýpková Z. (2003) Plot sizes used for phytosociological sampling of European vegetation. *Journal of Vegetation Science* **14**, 563–570.

Chytrý, M., Tichý, M. Holt, J. & Botta-Dukát Z. (2002) Determination of diagnostic species with statistical fidelity measures. *Journal of Vegetation Science* **13**, 79-90.

Cowardin, L. M., Carter, V., Golet, F. C. & LaRoe, E. T. (1979) *Classification of the wetlands and deepwater habitats of the United States.* U.S. Fish and Wildlife Service, Washington, D.C., USA.

Curtis, J.T. (1959) *Vegetation of Wisconsin*. University of Wisconsin Press, Madison, Wisconsin. 657 pp.

De Cáceres, M., Font, X. & Oliva, F. (2008) Assessing species diagnostic value in large data sets: a comparison between phi coefficient and Ochiai index. *Journal of Vegetation Science* **19**, 779-788.

De Cáceres, M., Font, X., Vicente, P. & Oliva, F. (2009) Numerical reproduction of traditional classifications and automated vegetation identification. *Journal of Vegetation Science* **20**, 620-628.

De Cáceres, M. & Legendre, L. (2009) Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**, 3566-3574.

De Cáceres, M., Font, X. & Oliva, F. (2010a) The management of vegetation classifications with fuzzy clustering. Journal of Vegetation Science **21**, 1138-1151.

De Cáceres, M., Legendre, P. & Moretti, M. (2010b) Improving indicator species analysis by combining groups of sites. *Oikos* **119**, 1674-1684.

De Cáceres, M. & Wiser, S. (2012) Towards consistency in vegetation classification. *Journal of Vegetation Science **23**, in press* (DOI: 10.1111/j.1654-1103.2011.01354).

Dengler, J. (2009) A flexible multi-scale approach for standardised recording of plant species richness patterns. *Ecological Indicators* **9**, 1169-1178.

Dengler, J., Chytrý, M. & Ewald, J. (2008) Phytosociology. *Encyclopedia of Ecology* (eds. S. E. Jørgensen & B. D. Fath), pp. 2767-2779. Elsevier, Oxford, UK.

Dengler, J., Jansen, F., Glöckler, F., Peet, R.K., De Cáceres, M., Chytrý,M., Ewald, J., Oldeland, J., Finckh, M., Mucina, M., Schaminée, J.H.J., & Spencer, S. (2011) The Global Index of Vegetation-Plot Databases: a new resource for vegetation science. *Journal of Vegetation Science* **22**, 582-597.

Dengler, J., Löbel, S. & Dolnik, C. (2009) Species constancy depends on plot size - a problem for vegetation classification and how it can be solved. *Journal of Vegetation Science* **20**, 754-766.

Domin, K. (1928) The relations of the Tatra mountain vegetation to the edaphic factors of the habitat: a synecological study. *Acta Botanica Bohemica* **6/7**, 133–164.

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

Ellenberg, H. (1956) *Grundlagen der Vegetationsgliederung. 1. Teil : Aufgaben und Methoden der Vegetationskunde.* In H. Walter, editor. Einfhrung in die Phytologie. Ulmer Stuttgart, 136 pp.

Equihua, M. (1990) Fuzzy clustering of ecological data. *Journal of Ecology* **78**, 519-534.

Ewald, J. (2003) A critique for phytosociology. *Journal of Vegetation Science* **14**, 291–296.

Faith.D.P., Minchin, P.R. & Belbin, L. (1987) Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69**, 57--68.

Feoli, E. & Orlóci, L. (1979) Analysis of concentration and detection of underlying factors in structured tables. *Vegetatio* **40**, 49-54.

Franz, N. M., Peet, R.K. & Weakley, A.S. (2008) On the use of taxonomic concepts in support of biodiversity research and taxonomy. Symposium Proceedings, In: Wheeler, Q. D., Ed., *The New Taxonomy*. Systematics Association Special Volume **74**, 63-86. Taylor & Francis, Boca Raton, FL.

Fridley, J.D., Peet, R.K. Wentworth, T.R. & White, P.S. (2005) Connecting fine- and broad-scale patterns of species diversity: species-area relationships of Southeastern U.S. flora. *Ecology* **86**, 1172-1177.

Gégout, J-C. & Coudin, C. (2012) The right relevé in the right vegetation unit: a new typicality index to reproduce expert judgment with an automatic classification programme. *Journal of Vegetation Science* **23**, *in press* (DOI: 10.1111/j.1654-1103.2011.01337.x).

Gillison, A.N. & Brewer, K.R.W. (1985) The use of gradient directed transects or gradsects in natural resource survey. *Journal of Environmental Management* **20**, 103-17.

Gleason, H.A. (1926) The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club* **53**, 7-26.

Gleason, H.A. (1939) The individualistic concept of the plant association. *American Midland Naturalist* **21**, 92-110.

Goodall, D.W. (1953) Objective methods for the classification of vegetation. II.Fidelity and indicator value. *Australian Journal of Botany* **1**, 434-456.

Goodall, D.W. (1973) Sample similarity and species correlation. Handbook of Vegetation Science, Dr. W. Junk, the Hague.

Goodman L. & Kruskal, W. (1954) Measures of association for cross-validations. *Journal of the American Statistical Association* **49**, 732-764.

Gopal, S. & Woodcock, C. (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing* **60**, 181-188.

Gray, A.N., Brandeis, T.J., Shaw, J.D. & McWilliams, W.H. (2012) Forest inventory vegetation database of the United States of America. *Biodiversity and Ecology*, *in press*.

Grossman, D. H., Faber-Langendoen, D., Weakley, A. S., Anderson, M., Bourgeron, P., R. Crawford, R., Goodin, K., Landaal, S., Metzler, K., Patterson, K.D., Pyne, M., Reid, M. & Sneddon, L. (1998) *International classification of ecological communities: terrestrial vegetation of the United States. Volume I. The National Vegetation Classification System: development, status, and applications.* The Nature Conservancy, Arlington, Virginia, USA.

Hartigan, J. A & Wong, M. A. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **28**, 100--108.

Hennekens, S.M. *&* Schaminée, J.H.J. (2001) TURBOVEG, a comprehensive data base management system for vegetation data*. Journal of Vegetation Science* **12***:* 589**–**591.

Hill, M.O. (1979) TWINSPAN — A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca, NY, US.

Hubálek, Z. (1982) Coefficients of association and similarity, based on binary (presence--absence) data: an evaluation. *Biological Review* **57**, 669-689.

Hubert, L.J. & Levin, J.R. (1976) A general framework for assessing categorical clustering in free recall. *Psychology Bulletin* **83**, 1072-1080.

Jansen, F. & Dengler, J. (2010) Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* **21**, 1179-1186.

Jennings, M.D., Faber-Langendoen, D., Loucks, O.L., Peet, R.K. & Roberts, D. (2009) Characterizing Associations and Alliances of the U.S. National Vegetation Classification. *Ecological Monographs* **79**, 173-199.

Jones, M.B., Schildhauer, M.P., Reichman, O.J. & Bowers, S. (2006) The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution and Systematics* **37**, 519–544.

Juhász-Nagy, P. (1964) Some theoretical models of cenological fidelity I. *Acta Botanica Debrecina* **3**, 33-43.

Kaufman, L. & Rousseeuw, P.J. (1990) *Finding groups in data.* John Wiley and Sons, New York.

Kent, M. (2012) *Vegetation description and data analysis: A practical approach*. Second edition. Wiley-Blackwell, Oxford, UK.

Knollová, I., Chytrý, M., Tichý, L. & Hájek, O. (2005) Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies. *Journal of Vegetation Science* **16**, 479-486.

Krajina, V. J. (1933) Die Pflanzengesellschaften de Mlynica-Tales in den Vysoke Tatry (Hohe Tatra). Mit besonderer Berücksichtigung der ökologischen Verhältnisse. *Beiheftezum Botanische Centralblatt* **50**, 774–957; **51**, 1–224.

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

Lance, G.N. & Williams, W.T. (1967) A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal* **9**, 373-380.

Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. Oecologia **129**, 271-280.

Legendre, P. & Legendre, L. (1998) *Numerical ecology, 2nd*. Edition. Developments in Environmental Modelling 20. Elsevier, Amsterdam, NL. 853 pp.

Lengyel, A., Chytrý, M. & Tichý, L. (2011) Heterogeneity-constrained random resampling of phytosociological databases. *Journal of Vegetation Science* **22**, 175-183.

Lepš, J. & Šmilauer, P. (2003) *Multivariate analysis of ecological data using CANOCO*. Cambridge University Press, Cambridge, UK. 269 pp.

Ludwig, J.A. & Reynolds, J.F. (1988) *Statistical ecology: A primer on methods and computing*. John Wiley and Sons, New York. 337 pp.

McCune, B. (1994) Improving community analysis with the Beals smoothing function. *Ecoscience* **1**, 82–86

McCune, B. & Grace, J.B. (2002) *Analysis of ecological communities*. MjM Software Design, Gleneden Beach, Oregon. 300 pp.

Michalcová, D., Lvončík, S., Chytrý, M. & Hájek, O. (2011) Bias in vegetation databases? A comparison of stratified-random and preferential sampling. *Journal of Vegetation Science* **22**, 281-291.

Mucina, L. (1997) Classification of vegetation: Past, present and future. *Journal of Vegetation Science* **8**, 751-760.

Mucina, L., Rodwell, J.S., J.H.J. Schaminée, J.H.J. & Dierschke, H. (1993) European vegetation survey: Current state of some national programs. *Journal of Vegetation Science* **4**, 429-438.

Mucina, L., Schaminée, J.H.J & Rodwell, J.S. (2000) Common data standards for recording relevés in field survey for vegetation classification. *Journal of Vegetation* Science **11**, 769-772.

Mueller-Dombois, D. & Ellenberg, H. (1974) *Aims and methods of vegetation ecology*. Wiley, New York. 547 pp.

Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography* **26**, 867-878.

Noest, V., van der Maarel, E. & Van der Muelen, F. (1989) Optimum transformation of of plant species cover-abundance values. *Vegetatio* **83**, 167-178.

Orlóci, L. (1967) An agglomerative method for classification of plant communities. *Journal of Ecology* **55**, 193-206.

Orlóci, L. (1978) Multivariate analysis in vegetation research. 2nd Edition, Dr. W. Junk, The Hague. 451 pp.

Peet, R.K., Lee, M.T., Jennings, M.D., & Faber-Langendoen, D. (2012) VegBank: A Permanent, Open-Access Archive for Vegetation Plot Data. *Biodiversity and Ecology, in press*.

Peet, R. K., Wentworth, T.R. & White, P.S. (1998) The North Carolina Vegetation Survey protocol: a flexible, multipurpose method for recording vegetation composition and structure. *Castanea* **63**, 262–274.

Pfister, R. D., Kovalchik, B. L., Arno, S. F. & Presby, R. C. (1977) Forest habitat types of Montana. USDA Forest Service Intermt. For. and Range Exp. Stn., Gen. Tech. Rep. INT-34; 174 pp.

Pfister, R.D. & Arno, S.F. (1980) Classifying forest habitat types based on potential climax vegetation. *Forest Science* **26**, 52-70.

Podani, J. (1990) Comparison of fuzzy classifications. *Coenoses* **5**, 17-21.

Podani, J. (2000) Simulation of random dendrograms and comparison tests: Some comments. *Journal of Classification* **17**, 123-142.

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

Podani, J. & Csányi, B. (2010) Detecting indicator species: Some extensions of the IndVal measure. *Ecological Indicators* **10**, 1119-1124.

Podani, J. & Feoli, E. (1991) A general strategy for the simultaneous classification of variables and objects in ecological data tables. *Journal of Vegetation Science* **2**, 435-444.

Radford, A.E., Ahles, H.E. & Bell, C.R. (1968) *Manual of the vascular flora of the Carolinas*. University of North Carolina Press, Chapel Hill, North Carolina, USA. 1183 pp.

Roberts, D.W. (2010) optpart: Optimal partitioning of similarity relations. R package version 2.0-1, http://CRAN.R-project.org/package=optpart.

Rodríguez, J.P., Rodríguez-Clark, K.M., Baille, J.E.M., Ash, N., Benson, J. Boucher, T., Brown, C., Burgess, N.D., Collen, B., Jennings, M., Keith, D.A., Nicholson, E., Revenga, C., Reyers, B., Rouget, M., Smith, T., Spalding, M., Taber, A., Walpole, M., Zager, I. & Zamin, T. (2011) Establishing IUCN redlist criteria for threatened ecosystems. *Conservation Biology* **25**, 21-29.

Rodwell, J.S. (2006) *National Vegetation Classification: User's Handbook.* Joint Nature Conservation Committee, Peterborough, UK. 68 pp.

Rodwell, J.S., Pignatti, S., Mucina, L. & Schaminée, J.H.J. (1995) European Vegetation Survey: update on progress. *Journal of Vegetation Science* **6**, 759-762.

Rodwell, J.S., Schaminée, J.H.J., Mucina, L., Pignatti, S., Dring, J. & Moss, D. (2002*) The diversity of European vegetation*. Wageningen. 168 pp.

Roleček, J., Chytrý, M., Háyek, M., Lvoncik, S. & Tichý, L. (2007) Sampling in large-scale vegetation studies: Do not sacrifice ecological thinking to statistical puritanism. *Folia Geobotanica* **42**, 199-208.

Roleček, J., Tichý, L., Zleney, D. & Chytrý, M. (2009) Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. *Journal of Vegetation Science* **20,** 596-602.

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computation and Applied Math*ematics **20**, 53-65.

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., Hennekens, S.M. & Ozinga, W.A. (2007) Use of the ecological information system SynBioSys for the analysis of large datasets. *Journal of Vegetation Science* **18**, 463-470.

Shimwell, D.W. (1971) *Description and classification of vegetation*. Sidgwick and Jackson, London, UK. 322 pp.

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.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry. The principles and practice of statistics in biological research, 3rd. edition*. Freeman, New York, NY, US.

Sokal, R.R. & Sneath, P.H.A. (1963) *Principles of numerical taxonomy*. Freeman. San Francisco and London. 359 pp.

Stohlgren, T. J., Falkner, M.B. & Schell, L.D. (1995) A modified-Whittaker nested vegetation sampling method. *Vegetatio* **117**, 113-121.

Szafer, W & Pawlowski, B (1927) Die Pflanzenassoziationen des Tatra-Gebirges. A. Bemerkungen fiber die angewandte Arbeitsmethodik. *Bull. Int. Acad. Pol. Sci. Lettr. Cl. Sci Nat. Math., B*, **1926** (12), Suppl.: 1-12.

TDWG (2005) Taxonomic Concept Transfer Schema. Biodiversity Information Standards. http://www.tdwg.org/standards/117/.

Tichý, L. & Chytrý, M. (2006) Statistical determination of diagnostic species for site groups of unequal size. *Journal of Vegetation Science* **17**, 809-818.

Tichý, L., Chytrý, M., Hájek, M., Talbot, S.S. & Botta-Dukát Z. (2010) OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science* **21**, 287-299.

Tsiripidis, I., Bergmeier, E., Fotiadis, G. & Dimopoulos, P. (2010) A new algorithm for the determination of differential taxa. *Journal of Vegetation Science* **20**, 233-240.

USFGDC (United States Federal Geographic Data Committee). (2008*) National Vegetation Classification Standard, Version 2 FGDC-STD-005-2008*. Vegetation Subcommittee, Federal Geographic Data Committee, FGDC Secretariat, U.S. Geological Survey. Reston, Virginia, USA.

van der Maarel, E. (1979) Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio* **39**, 97–144.

van der Maarel, E. (2007) Transformation of cover-abundance values for appropriate numerical treatment: Alternatives to the proposals by Podani. *Journal of Vegetation Science* **18**, 767-770.

van der Maarel, E. & Franklin, J. (2012) Vegetation ecology – an overview. Chapter 1, this volume.

van Tongeren, O., Gremmen, N. & Hennekens, S. (2008) Assignment of relevés by supervised clustering of plant communities using a new composite index. *Journal of Vegetation Science* **19**, 5255-536.

Vision, T.J. (2010) Open data and the social contract of scientific publishing. *BioScience* **60**, 330-331.

Waterton, C. (2002) From field to fantasy: Classifying nature, constructing Europe. *Social Studies of Science* **32**, 177-204.

Weber, H.E., Moravec, J. & Theurillat, J.-P. (2000) International Code of Phytosociological Nomenclature. 3rd edition. *Journal of Vegetation Science* **11**, 739-768.

Wesche, K. & von Wehrden, H. (2011) Surveying southern Mongolia: application of multivariate classification methods in drylands with low diversity and long floristic gradients. *Journal of Vegetation Science* **14**, 561-570.

Westhoff, V. & van der Maarel, E. (1973) The Braun-Blanquet approach. *Handbook of Vegetation Science* **5**, 617-726.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30**, 279-338.

Whittaker, R.H. (1962) Classification of natural communities. *Botanical Review* **28**, 1-239.

Whittaker, R. H., Niering, W.A. & Crisp, M.D. (1979) Structure, pattern, and diversity of a mallee community in New South Wales. *Vegetatio* **39**, 65-76.

Whittaker, R.H., editor. (1973) *Handbook of Vegetation Science – Part V. Ordination and classification of communities.* Junk, The Hague. 737 pp.

Wildi, O. (2010) *Data analysis in vegetation ecology.* Wiley-Blackwell, Oxford, UK. 211 pp.

Williams, W.T., Lambert, J.M & Lance, G.N. (1966) Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. *Journal of Ecology* **54**, 427-445.

Willner, W., Tichý, L. & Chýtrý, M. (2009) Effects of different fidelity measures and contexts on the determination of diagnostic species. *Journal of Vegetation Science* **20**, 10-137.

Wilson, J.B. (*submitted*) There is no need to estimate species abundance in community surveys, presence/absence is actually better.

Wiser, S., Spencer, N., De Cáceres, M., Kleikamp, M., Boyle, B., & Peet, R.K. (*2011*) Veg-X – An exchange standard for plot-based vegetation data. *Journal of Vegetation Science* **22**, 598-609.

Table 4.1. A comparison of several cover scales used for recording vegetation plots including the traditional Braun-Blanquet scale (1928), the original Domin scale (1928), a variant of the Domin scale by Krajina (1933), and the scales of the Carolina (Peet *et al.* 1998) and New Zealand vegetation surveys (Allen 1992). The shading indicates how the newer indices next into the Braun-Blanquet scale.

Range of cover	Braun-Blanquet	*Domin*	Krajina	Carolina	New Zealand
Single individual	r	+	+	1	1
Sporadic or few	+	1	1	1	1
0-1%	1	2	1	2	1
1-2%	1	3	1	3	2
2-3%	1	3	1	4	2
3-5%	1	4	1	4	2
5-10%	2	4	4	5	3
10-25%	2	5	5	6	3
25-33%	3	6	6	7	4
33-50%	3	7	7	7	4
50-75%	4	8	8	8	5
75-90%	5	9	9	9	6
90-95%	5	10	9	9	6
95-100%	5	10	10	10	6

Table 4.2 Definitions of Dissimilarity and Distance. d_{ij} = the dissimilarity of plot *I* to plot *j*, x_{ik} = the abundance of species *k* in plot *I* for *p* species, x_{i+} = the sum of abundances for all species in plot *I*.

Index	Equation
Bray-Curtis	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert \Big/ \sum_{k=1}^{p} x_{ik} + x_{jk}$
Marczewski-Steinhaus	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert \Big/ \sum_{k=1}^{p} max(x_{ik}, x_{jk})$
Euclidean Distance	$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
Manhattan Distance	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
Hellinger Distance	$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(\sqrt{\frac{x_{ik}}{x_{i+}}} - \sqrt{\frac{x_{jk}}{x_{j+}}} \right)^2}$
Chord Distance	$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(\frac{x_{ik}}{\sqrt{x_{i+}^2}} - \frac{x_{jk}}{\sqrt{x_{j+}^2}} \right)^2}$

Table 4.3 Hierarchical Agglomerative Clustering Criteria. d_{AB} = dissimilarity between cluster A and B, d_{ij} = the dissimilarity between plots *i* and *j*, $i \in A$ = plot *i* is a member of set A, $|A|$ = the number of members of cluster A, $\overline{d_A}$ = the mean coordinate on axis d for plots in cluster A, and Var d_k = the variance of dissimilarities formed in fusing cluster A with B.

Linkage	Equation
Single	$d_{A,B} = min\{d_{ij} : i \in A, j \in B\}$
Complete	$d_{A,B} = max\{d_{ij} : i \in A, j \in B\}$
Average	$d_{A,B} = \frac{1}{\|A\| \times \|B\|} \sum_{i \in A} \sum_{j \in B} d_{ij}$
Centroid	$d_{A,B} = \sqrt{\sum_{d=1}^{D} (\overline{d_A} - \overline{d_B})^2}$
Ward's	$d_{A,B} = Var\ d_k : k \in A \cup B$

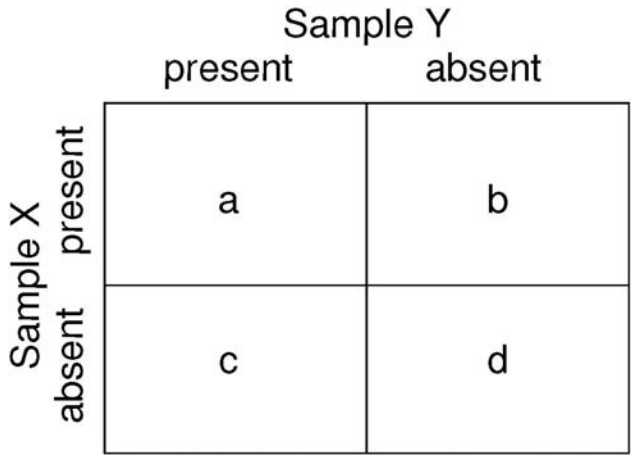

Fig. 4.1 Contingency table notation for presence/absence dissimilarity indices.

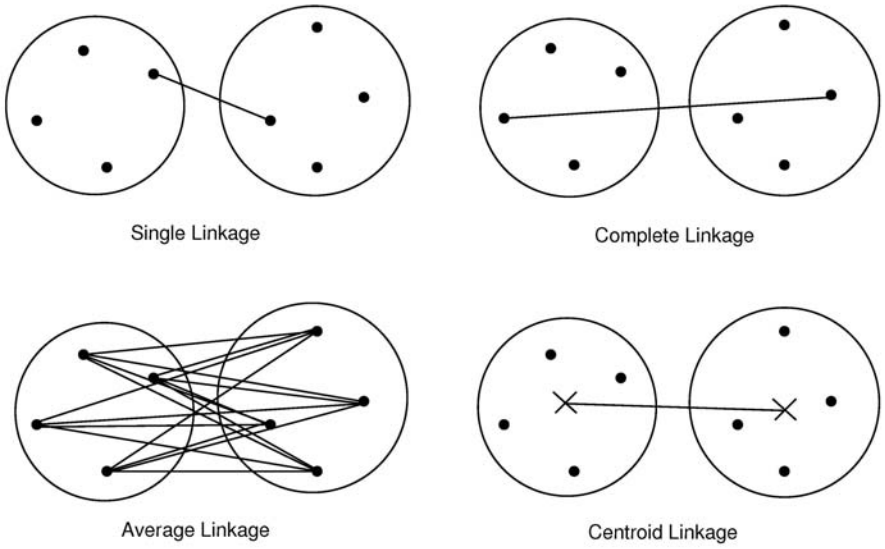

Fig 4.2 Hierarchical agglomerative algorithms differ specifically in how they define dissimilarity between clusters with more than one member.

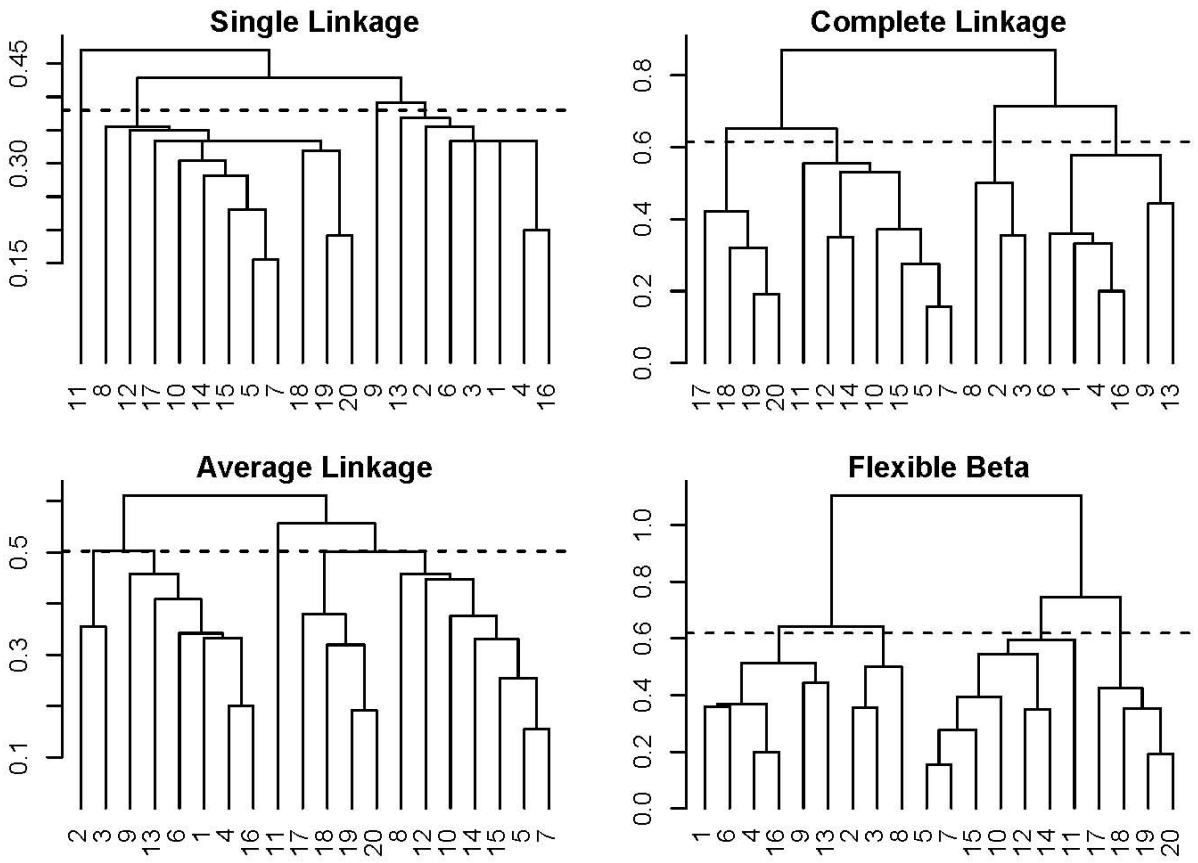

Fig 4.3 Dendrograms for hierarchical agglomerative clustering algorithms based on the same dissimilarity matrix but using different linkages