Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 1078–1083 doi: 10.1111/j.2041-210X.2012.00232.x

APPLICATION

TAXONSTAND: An R package for species names standardisation in vegetation databases

Luis Cayuela[1]*, Íñigo Granzow-de la Cerda[2], Fabio S. Albuquerque[3] and Duncan J. Golicher[4]

[1]Departamento de Biología y Geología, Universidad Rey Juan Carlos, c/ Tulipán s/n, E-28933 Móstoles, Madrid, Spain; [2]Departamento de Biología Animal, Biología Vegetal y Ecología, Universidad Autónoma de Barcelona, E-08193 Barcelona, Spain; [3]Departamento de Ecología, Centro Andaluz de Medio Ambiente, Universidad de Granada, E-18006 Granada, Spain; and [4]Centre for Conservation, Ecology & Environmental Change, School of Conservation Sciences, Dorset House DG 38B, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK

Summary

1. Compilation of vegetation databases has contributed significantly to the advancement of vegetation science all over the world. Yet, methodological problems result from the use of plant names, particularly in data that originate from numerous and heterogeneous sources. One of the main problems is the inordinate number of synonyms that can be found in vegetation lists.

2. We present Taxonstand, an R package to automatically standardise plant names using The Plant List (http://www.theplantlist.org). The scripts included in this package allow connection to the online search engine of the Plant List and retrieve information from each species about its current taxonomic status. In those cases where the species name is a synonym, it is replaced by the current accepted name. In addition, this package can help correcting orthographic errors in specific epithets.

3. This tool greatly facilitates the preparation of large vegetation databases prior to their analyses, particularly when they cover broad geographical areas (supranational or even continental scale) or contain data from regions with rich floras where taxonomic problems have not been resolved for many of their taxa. Automated workflows such as the one provided by the taxonstand package can ease considerably this task using a widely accessible working nomenclatural authority list for plant species names such as The Plant List.

Keywords: data cleaning, nomenclature, taxonomy, The Plant List.

Introduction

Vegetation databases for species local occurrence and species checklists are compiled all over the globe. Several initiatives have emerged since the beginning of the 21st century to acquire, manage and provide access to information in research collections such as natural history museums and herbaria, as well as data gathered by observational and survey projects held by universities, government research centres, nongovernmental organisations and private institutions and individuals. Possibly, the most inclusive of these initiatives is the Global Biodiversity Information Facility (GBIF, http://www.gbif.org), which promotes and facilitates the access, discovery and use of information about the occurrence of taxa across the planet (Yesson *et al.* 2007). For plants, several research networks at regional or continental scales have also emerged during the last decades, which have contributed to the creation and management of several vegetation databases (see Dengler *et al.* 2011 and references herein). All these initiatives and networks have contributed significantly to the advancement of vegetation science, opening new venues for integrated data analysis at different scales, including, amongst others, predictive mapping, detection of hotspots and conservation prioritisation, classification of vegetation types, and tests for fundamental ecological hypotheses regarding functional traits, assembly rules and biodiversity patterns (Dengler *et al.* 2011).

Despite the potential of vegetation databases, methodological problems resulting from the heterogeneity of data sources still hinder their application to research questions (Ewald 2003; Jansen & Dengler 2010). Common problems include geographical biases – for example, records that are highly

*Correspondence author. E-mail: luis.cayuela@urjc.es correlated spatially with road or river networks (Cayuela *et al.* 2009 and references herein), spatial errors in geo-referenced records (Chapman 2005) and problems associated with taxonomic concepts, including the use of plant names (Jansen & Dengler 2010). Whereas geographical biases and spatial errors in vegetation databases have received much attention in the scientific literature (Neldner, Crossley & Cofinas 1995; Hortal, Lobo & Jiménez-Valverde 2007; Cayuela *et al.* 2009) and partial solutions have been proposed for handling these problems (e.g. Stockwell & Peterson 2002; Knollová *et al.* 2005), little attention has been given to nomenclatural and – to some extent – taxonomic problems of plant names in vegetation databases (but see Jansen & Dengler 2010). In addition to inflating diversity overall, the overwhelming number of synonyms that plague the botanical literature is also responsible, if not addressed, for generating all sorts of errors when analysing taxonomic assemblages for a given region. At least *c.* 46% of all species names are estimated to be synonyms (figure from The Plant List project, http://www.theplantlist.org/statistics). A synonym, as defined in the International Code of Botanical Nomenclature (McNeill *et al.* 2006), is a name considered to apply to the same taxon as the accepted name. They may have two different origins: heterotypic synonym (taxonomic synonym), a synonym that is based on a type different from that of the accepted name; and homotypic synonym (nomenclatural synonym), a synonym that is based on the same type as that of another name in the same rank (from the ICBN). Plant checklists and vegetation databases are certainly not spared from this problem. Identification and correction of nomenclatural and taxonomic errors is also a critical step prior to conducting data analyses (Chapman 2005), and some important efforts have been made in this regard to automate procedures for cleaning data (Boyle 2006; TNRS 2012).

In this paper, we present TAXONSTAND, an R-based package to automatically standardise plant names using a universally accessible authority table. We have adopted The Plant List (http://www.theplantlist.org), a comprehensive broadly accepted and widely accessible working list of known plant species that has been developed and disseminated in direct response to the Global Strategy for Plant Conservation (http://www.cbd.int/gspc/), adopted in 2002 by the 193 governments that are signatories of the Convention on Biological Diversity. The Plant List was produced as a collaborative effort coordinated by the Royal Botanic Gardens, Kew, and the Missouri Botanical Garden, with the involvement of a number of collaborating entities worldwide. Although it is not perfect and represents work in progress, it is currently the most comprehensive authority list for plant names (Kalwij 2012). Therefore, it is increasingly used in international initiatives (e.g. BIOTREE-NET, Cayuela *et al.* 2011) as a reference source for resolving or verifying the spelling of plant names and a means to find from a global view the botanically accepted name for a plant and all of its alternative synonyms. The TAXONSTAND package allows connection to the online search engine of The Plant List. Users need to provide either a single scientific plant name or a list of species names, and the package searches for the corresponding accepted combination for each name provided. In those cases where the species name provided as input is recognised as a synonym, TAXONSTAND finds the current accepted name and returns both the original and the accepted name. The package also incorporates an approximate string matching algorithm that allows recognition of orthographic errors in names in the input set. The output provides the full taxonomic name accepted (unless unresolved), with authority, and familial circumscription as recognised in The Plant List. Overall, we aim to provide a functional interface to The Plant List website through the widely used R environment to help plant scientists – often not familiar with the intricacies of botanical nomenclature – to rigorously, rapidly and at no cost, overcome some of the most common methodological problems associated with the use of plant names in vegetation databases, particularly those compiled from heterogeneous data sources.

Problems associated with the use of plant names

Plant taxonomy is a dynamic discipline (Stuessy 2009). A consequence of this constant taxonomic flux and individual idiosyncrasy of taxonomic understanding is that the thousands of floras and checklists in use worldwide are seldom congruent in their taxonomy and nomenclature. This becomes a real problem when data that originate from various and heterogeneous sources are put together for analyses.

One hurdle in compiling unified, entirely agreed-upon vegetation lists is the ever-changing systematic (phylogenetic) interpretation of taxonomic relationships. The result is multiple combinations for most taxonomic entities, whether it is attributable to nomenclatural or to taxonomic reasons: taxa are lumped or split, species are subordinated to others or are moved from one genus to another in response to – far from universal – conflicting taxonomic criteria. In other words, species – to the dismay of nontaxonomists – change names (often back and forth) at an exceedingly high pace because of reconsideration of generic and specific concepts, or in some cases to correct nomenclatural mistakes. The result is an inordinate number of synonyms for almost every species. The use of vegetation databases where synonyms have not yet been identified as such and sorted out will greatly overestimate biodiversity at different scales (Isaac, Mallet & Mace 2004) and may contribute to an inaccurate delimitation of species distribution ranges (Jansen & Dengler 2010). Often, owing to a broad consensus on the taxonomic circumscription for a certain taxon, there is general agreement for accepting a given species name and regarding the rest as (nonaccepted) synonyms. However, publications that precede the shift that led to the consensus will contain 'old' names – now disfavoured – although this may not be necessarily known to the nonspecialist producing or analysing sizeable vegetation databases. Therefore, a mechanism needs to be put in place for identifying fairly broadly recognised reductions to synonymy. In fact, it is commonplace to find many plant name lists with names now unaccepted that are regarded as synonyms. This is especially true for the less recent works and in those from regions with rich floras where taxonomic problems have not been resolved for many of their taxa.

Taxonomic homonyms – albeit much less common than synonyms – can sometimes be found in vegetation databases. A homonym, as defined by the ICBN (McNeill *et al.* 2006), is a name spelled exactly like another name published for a taxon of the same rank based on a different type. The unlikely presence of homonyms only constitutes a problem in the unfortunate circumstances where the authority for names is not recorded in the database. This problem is hard to solve in practice in the presence solely of a list of plant names. If researchers have aprioristic knowledge on the distribution range of the different taxa, homonyms can be told apart on a case-by-case basis.

Of course, there is always the problem of misidentification, but another hurdle is the way taxonomic concepts are applied, especially at the infraspecific level (Jansen & Dengler 2010). Neither of these particulars are easy to resolve, although the latter can be addressed with some success through the 'taxon view' schema discussed by Jansen & Dengler (2010), where recording both the original plant name along with a link to a reference that defines the taxon concept. The issue of taxonomic circumscription or concept is certainly there, but not the most serious. It is very unfortunate that many vegetation studies do not observe the basic principle of providing taxa names with an authority. Jansen & Dengler (2010) make a strong case of this prevalent oversight, pointing out that authority names will resolve a good deal of ambiguity for recognising homonyms. However, the nature of vegetation inventories, where a great deal of taxa are involved, often requires overlooking some of the more precise taxonomic criteria, a type of precision that is more appropriate for taxonomic revisions. And whilst this may be a good contribution for future collection efforts, it is overall impractical for most vegetation databases. As pointed out by Schaminée *et al.* (2009), as much as 60% of the digitally available data sets for vegetation plots in Europe are stored without a taxon view link, and, in our experience, this figure may be even higher in tropical vegetation databases (e.g. Cayuela *et al.* 2011).

Of the problems that are most common when analysing vegetation lists, only the occurrence of synonyms can be adequately addressed using The Plant List through our R package TAXONSTAND. Whilst the R package VEGDATA (Jansen & Dengler 2010) provides appropriate tools to avoid flawed results that would ensue from inconsistent use of plant names, it does require the presence of a taxonomic authority table in the form of a checklist – often region specific – which is certainly not available in many situations, particularly for regions with rich floras such as the tropics. Unfortunately, automated recognition of taxonomic homonyms is not yet resolved if name authorities are not provided. However, incidence of homonyms in databases is relatively low and constitutes a minor problem, at least quantitatively.

Workflow for taxonomic standardisation in *TAXONSTAND*

Two functions are available within the TAXONSTAND package (version 1.0, available at http://cran.r-project.org/web/packages/Taxonstand/): 'TPLck' and 'TPL'. 'TPLck' connects to The Plant List and validates the name of a single plant species name, replacing synonyms for accepted names and correcting orthographic errors in target plant names. Function 'TPL' applies function 'TPLck' to a list of species names. Overall, these two functions together perform four basic actions:

1. Nomenclatural standardisation. TAXONSTAND standardises all species names using The Plant List database, replaces synonyms with the current accepted names (as recognised in The Plant List) and stores the original name in a separate field. The nomenclatural status for each name as *accepted*, a *synonym* or *unresolved* is returned in a dedicated field.

2. Recognition and removal of standard annotations. Identification qualifiers such as 'cf.', 'aff.', 's.l.', and 's.str.' and their orthographic variants are removed and stored in a separate field.

3. Recognition and correction of orthographic errors in specific epithets.

4. Return of the family name to which the taxon belongs and authority names.

The workflow is designed to conduct a sequence of concatenated steps summarised in Fig. 1. First, if the genus is not recognised, no output is returned from The Plant List website. This can be due to orthographic errors in the genus name or simply because it has not been yet incorporated into The Plant List. Orthographic errors in the genus name cannot be identified nor corrected by the TAXONSTAND schema.

If the genus name is found in The Plant List but without specific epithet, The Plant List returns the entire list of species in the genus. The schema runs an approximate string match between the target's specific epithet and the list of species produced by The Plant List. If a match is found, the epithet is treated as an orthographic error and corrected, and The Plant List is queried again for the complete corrected name. If it fails to find a match, it can be due to the same reasons as stated before at the genus level. Note that a common feature of lists of species names, particularly when obtained from field inventories in highly diverse regions, is the use of morphospecies, that is, species whose identity is based on morphological features for recognition and differentiation from congeners, but that do not correspond to a valid name. It is important that species names are appropriately labelled, for example, using 'sp' or 'sp.' in the specific epithet, to avoid false positives in the fuzzy match.

If both genus name and specific epithet are found in The Plant List, two scenarios can occur. First, only one match is found: if the name is regarded as accepted or unresolved, no action is taken and the name is returned unchanged; if the target name is regarded by The Plant List as a synonym, the corresponding accepted name is retrieved to replace the target name. Second, when more than one name is returned by The Plant List (e.g. various subordinate infraspecific taxa), TAXONSTAND searches for matches at the infraspecific level: if no match is found, it retrieves the first accepted specific epithet, disregarding infraspecific taxa; if no accepted names are found, *TAXOSTAND* searches for the first synonym, ignoring infraspecific, and retrieves the accepted name for that synonym to

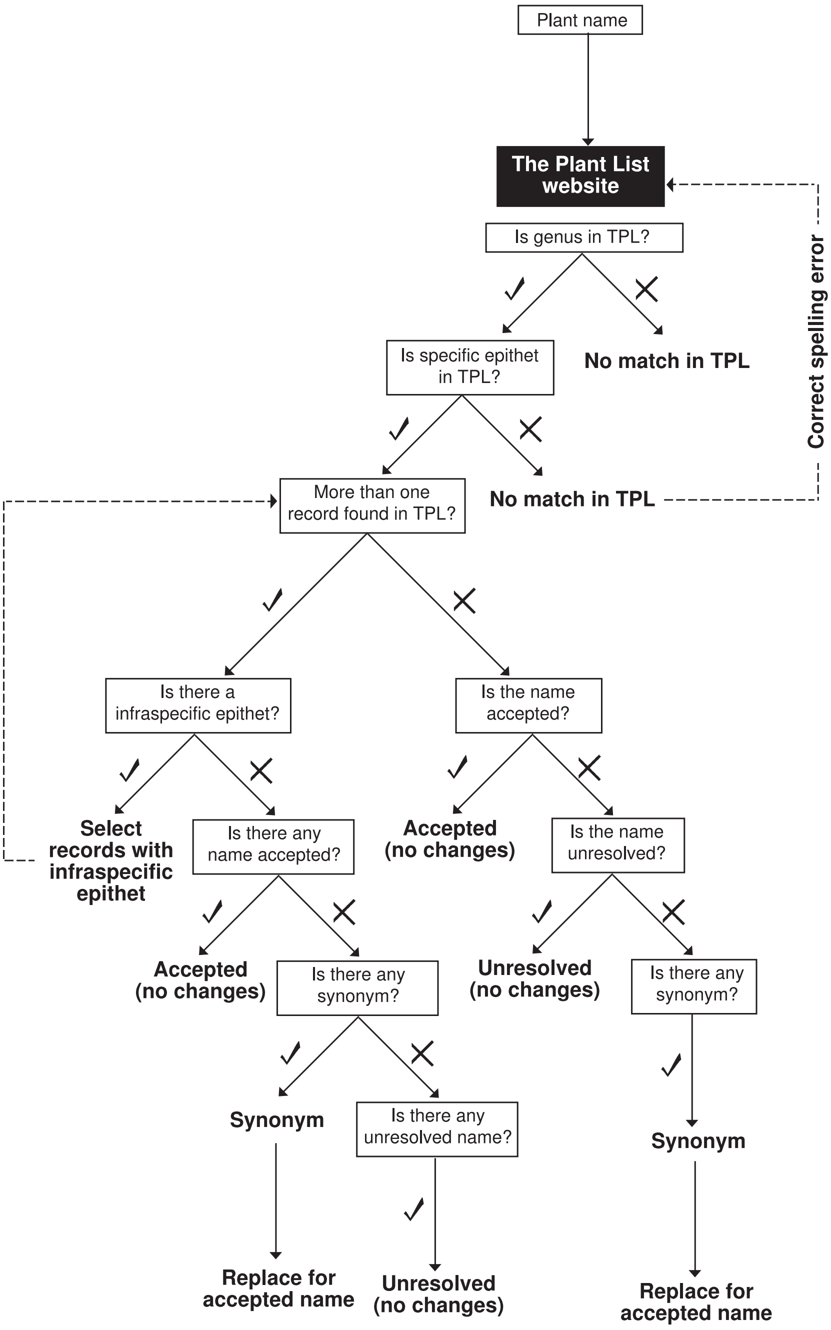

Fig. 1. Schematic view of the workflow for taxonomic standardisation by TAXONSTAND. The process involves various concatenated steps (white boxes), which apply to each species name. The output of each step will determine each subsequent step until a solution is reached (bold text).

replace the target; in all other cases, TAXOSTAND returns the first unresolved name, ignoring subspecies and varieties. If all the names found in the search correspond to infraspecific taxa, whilst there is no infraspecific recognition for the target name, then the first name in the list is chosen by default. The Plant List displays taxa alphabetically by epithet, regardless of infraspecific rank.

The output of a function 'TPL' run on a target list of species names returns the following fields: (i) Genus – genus name from target; (ii) Species – species epithet from target; (iii) Abbrev – Standard annotation used in species epithet, including 'cf.', 'aff.', 's.l.' and 's.str.' and their orthographic variants; (iv) Infraspecific – infraspecific epithet from target. If 'infra = FALSE' not shown; (v) Plant.Name. Index – logical. If 'TRUE', the name is in The Plant List; (vi) Taxonomic.status – 'Accepted', 'Synonym' or 'Unresolved'; (vii) Family – family name as recognised by The Plant List; (viii) New.Genus – accepted or unresolved genus name from The Plant List; (ix) New.Species – accepted or unresolved specific epithet from The Plant List; (x) New.Infraspecific – accepted or unresolved infraspecific epithet from The Plant List; (xi) Authority – as returned by The Plant List for the accepted or unresolved name; (xii) Typo – logical. If 'TRUE', there was an orthographic error in the specific epithet that has been corrected; (xiii) WFormat – logical. If 'TRUE', data in The Plant List are the wrong format (not properly tabulated) and data cannot be retrieved automatically.

At present, the package has been tested with several plant names lists, including (i) a list of bryophytes from 26 Mediterranean islands containing 1122 taxon names (compiled by I. Granzow-de la Cerda from the literature, unpublished results); (ii) a list of plant names for the region of Valencia, Spain, containing 3047 plant names (J. Tormo, data retrieved from Banco de Datos de Biodiversidad, Comunidad Valenciana, available at http://bdb.cma.gva.es); (iii) the 'Amazonia' data set from the BETAPER package in R (Cayuela & de la Cruz 2009), containing 1188 tree scientific names (Higgins & Ruokolainen 2004); and (iv) the BIOTREE-NET data set containing 5113 tree species names (Cayuela *et al.* 2011). Overall, more than 10,000 plant names, including morphospecies, were checked against the Plant List. This task took approximately 2 h, although throughput is expected to vary because of computer hardware and speed of Internet connection. The results can be inspected in Table 1. The number of accepted names found in The Plant List ranged from 39·6% to 69·9%; the number of synonyms ranged from 3·1% to 25·6%; and the number of unresolved names ranged from 0·9% to 7·0%. It is noteworthy the large amount of nonavailable names in (iii) and (iv). This was mostly due to the presence of morphospecies in these data sets, which do not correspond to valid names. The amount of orthographic errors corrected by the fuzzy matching algorithm ranged from 0·7% to 5·8%. As a result of the standardisation, the number of names was reduced in all cases by *ca.* 3·5% in (2) to *c.* 15% in (i).

Contributions of *TAXONSTAND*

There are other applications that can review and correct names in species lists, such as Salvia's Taxonscrubber (Boyle 2006) or the Taxonomic Names Resolution Services (TNRS 2012). Whereas the former will only check orthographic issues and assign taxa to the appropriate family, without identifying synonyms, TNRS, in addition to correcting spelling errors and providing with alternative spellings from a standard list of names, it converts synonyms to the accepted name. Like TAXONSTAND, TNRS can process many names at a fast pace, saving hours of tedious and error-prone manual name correction. For names that cannot be resolved automatically, TNRS presents, as part of its workflow, alternative names and provides tools for researching and selecting the preferred name. The data source for TNRS is Tropicos (the Missouri Botanical Garden database, http://www.tropicos.org/), which is also one of the main data sources for The Plant List, although the latter draws largely from additional sources, namely WCSP (World Checklist of Selected Plant Families) for taxon names and IPNI for nomenclatural criteria. Therefore, a search conducted through TAXONSTAND will result in hits for more names that what TNRS would. To illustrate this point, we conducted a taxonomic standardisation with TNRS on one of the lists used for testing TAXONSTAND (see Table 1), namely the list of Mediterranean plant names from the region of Valencia, Spain. In contrast with the 2130 accepted names (69·9%) and 444 synonyms (14·6%) resulting from the use of TAXONSTAND, TNRS retrieved 1574 accepted names (51·6%), 459 synonyms (15·0%) and 832 names with no opinion (27·3%).

Finally, one of the strengths of TNRS is that it provides a user-friendly interface. This can be an attractive feature for most users. TAXONSTAND, on the other hand, requires operating in the R environment, which is not necessarily user-friendly, but provides by far more flexibility than a stand-alone web application like TNRS. Users can access and modify the source code for the functions included in the TAXONSTAND package to modify the input, search algorithms and/or the output as desired.

Conclusions

Vegetation databases have provided new venues for the exchange of data and have contributed significantly to the advancement of vegetation science. However, methodological problems resulting from the use of data originated from heterogeneous, nonstandardised, nomenclaturaly inconsistent sources hinder their usefulness in addressing research questions. Of these problems, those associated with the use of plant names have received particularly little attention to this date (but see Jansen & Dengler 2010). Synonyms are of particular importance since they are widespread in vegetation lists, will grossly overestimate biodiversity (Isaac, Mallet & Mace 2004), and contribute to the inaccurate delimitation of species

Table 1. Summary of output from the TAXONSTAND package on different lists of species names, including (i) a list of bryophytes from 26 Mediterranean islands (compiled by I. Granzow-de la Cerda from the literature, unpublished results); (ii) a list of plant names for the region of Valencia, Spain (J. Tormo, data retrieved from Banco de Datos de Biodiversidad, Comunidad Valenciana, available at http://bdb.cma.gva.es); (iii) the 'Amazonia' data set (Higgins & Ruokolainen 2004) from the BETAPER package in R (Cayuela & de la Cruz 2009); and (iv) the BIOTREE-NET data set (Cayuela *et al.* 2011). Accepted, synonyms, unresolved and nonavailable names add up to the total number of original names.

	Bryophytes	Mediterranean plants	Amazonian trees	Central American trees
Original names	1122	3047	1188	5113
Accepted	717 (63·9%)	2130 (69·9%)	471 (39·6%)	2595 (50·7%)
Synonym	287 (25·6%)	444 (14·6%)	37 (3·1%)	571 (11·2%)
Unresolved	79 (7·0%)	204 (6·7%)	11 (0·9%)	85 (1·7%)
Nonavailable	39 (3·5%)	269 (8·8%)	669 (56·3%)	1862 (36·4%)
Orthographic errors	17 (1·5%)	74 (2·4%)	8 (0·7%)	299 (5·8%)
Standardised names	955 (85·1%)	2940 (96·5%)	1040 (87·5%)	4720 (92·3%)

distribution ranges (Jansen & Dengler 2010). Standardisation of taxonomic names in plant databases that expand over broad geographical areas (supranational or even continental scales) can be a conflictive task. The use of a widely accessible working nomenclatural authority list for plant species names such as The Plant List can be useful as a standard to address the problem of redundancy in plant names that is generated by the presence of synonyms in various vegetation databases (Kalwij 2012). Yet, taxonomic standardisation can be quite time-consuming when compiling a large number of taxa names. Automated workflows such as the one provided by the TAXONSTAND package can ease considerably this task. In addition, this package can help correcting orthographic errors in specific epithets.

Acknowledgements

This study has been funded by Fundación BBVA as part of the project BIOTREE-NET (BIOCON08_044). Thanks to Jaume Tormo, Neptalí Ramírez-Marcial and Marcelino de la Cruz for testing a beta version of TAXONSTAND, providing insightful comments on its functionality, and suggesting further improvements.

References

Boyle, B.L. (2006) *TaxonScrubber, version 2.0. The SALVIAS Project.* http://www.salvias.net/pages/taxonscrubber.html [accessed 12 March 2012].

Cayuela, L. & de la Cruz, M. (2009) *betaper: functions to incorporate taxonomic uncertainty on multivariate analyses of ecological data. R package version 1.1-0.* http://CRAN.R-project.org/package = betaper.

Cayuela, L., Golicher, J.D., Newton, A.C., Kolb, M., Albuquerque, F.S., Arets, E.J.M.M., Alkemade, J.R.M. & Pérez, A.M. (2009) Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science*, **2**, 319–352.

Cayuela, L., Gálvez-Bravo, L., Pérez Pérez, R., Albuquerque, F.S., Golicher, D.J., Zahawi, R.A., Ramírez-Marcial, N., Garibaldi, C., Field, R., Rey Benayas, J.M., González-Espinosa, M., Balvanera, P., Castillo, M.A., Figueroa-Rangel, B.L., Griffith, D.M., Islebe, G.A., Kelly, D.L., Olvera-Vargas, M., Schnitzer, S.A., Velázquez, E., Williams-Linera, G., Brewer, S.W., Camacho-Cruz, A., Coronado, I., de Jong, B., del Castillo, R., Granzow-de la Cerda, I., Fernández, J., Fonseca, W., Galindo-Jaimes, L., Gillespie, T.W., González-Rivas, B., Gordon, J.E., Hurtado, J., Linares, J., Letcher, S.G., Mangan, S.A., Meave, J.A., Méndez, E.V., Meza, V., Ochoa-Gaona, S., Peterson, C.J., Ruiz-Gutierrez, V., Snarr, K.A., Tun Dzul, F., Valdez-Hernández, M., Viergever, K.M., White, D.A., Williams, J.N., Bonet, F.J. & Zamora, R. (2011) The Tree Biodiversity Network (BIOTREE-NET): prospects for biodiversity research and conservation in the Neotropics. *Biodiversity and Ecology*, **4**, in press.

Chapman, A.D. (2005) *Principles and methods of data cleaning – primary species and species-occurrence data, version 1.0.* Report for the Global Biodiversity Information Facility, Copenhagen.

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

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

Higgins, M.A. & Ruokolainen, K. (2004) Rapid tropical forest inventory: a comparison of techniques based on inventory data from western Amazonia. *Conservation Biology*, **18**, 799–811.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). *Conservation Biology*, **21**, 853–863.

Isaac, N.J.B., Mallet, J. & Mace, G.M. (2004) Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology and Evolution*, **19**, 464–469.

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

Kalwij, J.M. (2012) Review of 'The Plant List, a working list of all plant species'. *Journal of Vegetation Science*, **23**, in press.

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.

McNeill, J., Barrie, F.R., Burdet, H.M., Demoulin, V., Hawksworth, D.L., Marhold, K., Nicolson, D.H., Prado, J., Silva, P.C., Skog, J.E., Wiersema, J.H. & Turland, N.J. (eds) (2006) *International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventeenth International Botanical Congress. Vienna, Austria, July 2005.* Gantner Verlag, Ruggell, Liechtenstein. [Regnum Vegetabile 146. A.R.G. Gantner Verlag KG.].

Neldner, V.J., Crossley, D.C. & Cofinas, M. (1995) Using Geographic Information Systems (GIS) to determine the adequacy of sampling in vegetation surveys. *Biological Conservation*, **73**, 1–17.

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.

Stockwell, D.R.B. & Peterson, A.T. (2002) Controlling bias during predictive modeling with museum data. *Predicting species occurrences: Issues of scale and accuracy*. (eds J.M. Scott, P.J. Heglund, M. Morrison, M. Raphael, J. Haufler, B. Wall & F. Samson), pp. 537–546. Island Press, Covelo, California, USA.

Stuessy, T.F. (2009) *Plant Taxonomy: The Systematic Evaluation of Comparative Data*, 2nd edn. Columbia University Press, New York, NY, US.

TNRS, version 2.0. (2012) *The Taxonomic Name Resolution Service, taxonomic source: TROPICOS.* http://tnrs.iplantcollaborative.org/ [accessed 21 May 2012].

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. (2007) How Global Is the Global Biodiversity Information Facility? *PLoS One*, **2**, e1124.

Received: 30 March 2012; accepted 2 June 2012
Handling Editor: Robert Freckleton