Received: 11 June 2020 | Revised: 17 December 2020 | Accepted: 22 December 2020

DOI: 10.1111/avsc.12562

METHODS IN VEGETATION SCIENCE

Applied Vegetation Science IAVS

Implementing the formal language of the vegetation classification expert systems (ESy) in the statistical computing environment R

Helge Bruelheide[1,2] | Lubomír Tichý[3] | Milan Chytrý[3] | Florian Jansen[4]

[1]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany

[2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[3]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic

[4]Faculty of Agricultural and Environmental Sciences, Rostock University, Rostock, Germany

Correspondence
Helge Bruelheide, Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle, Germany.
Email: helge.bruelheide@botanik.uni-halle.de

Funding information
LT and MC were supported by the Czech Science Foundation (project no. 19-28491X).

Co-ordinating Editor: Sebastian Schmidtlein

Abstract

Aims: The machine-readable formal language of classification expert systems has become a standard for applying plot assignment rules in vegetation classification. Here we present an efficient algorithm implementing the vegetation classification expert system in the statistical programming language R.

Methods: The principal idea of the R implementation is to solve the assignments to vegetation types not sequentially plot by plot but to parse the assignment rules into (nested) components that each can be evaluated by simultaneous vector-based processing of all plots in a database.

Results and conclusions: We demonstrate the algorithm taking the EUNIS classification expert system of European habitat types (EUNIS-ESy) as an example. The R code version of the vegetation classification expert system is particularly useful in large vegetation-plot databases because it solves all logical operations vector-wise across all plots, allowing for efficient evaluation of membership expressions and formulas. Another advantage of the R implementation is that membership formulas are not only readable but can also be produced as a machine-written result, for example as the output of classification algorithms run in R.

KEYWORDS

COCKTAIL method, EUNIS, Europe, expert system, R software, vegetation classification, vegetation database

1 | INTRODUCTION

Consistent vegetation classifications have to be based on formalized rules (De Cáceres & Wiser, 2012). Since formalized vegetation has been proposed in the late 1990s (Bruelheide, 1997; Bruelheide & Jandt, 1997), applications of formalized classification to data sets of vegetation plots have steadily increased (see https://www.sci.muni.cz/botany/juice/?idm=25). A highly flexible and fully machine-readable expert system language for vegetation classification (ESy) has been developed by two of the authors of this paper (LT and MC) in several steps since the early 2000s (Kočí et al., 2003; Chytrý, 2007–2013; Landucci et al., 2015; Tichý et al., 2019). So far, the software tools for classifying vegetation using this language have been implemented in JUICE (Tichý, 2002) and TURBOVEG 3 (Hennekens, 2015). However, to allow a wider application and the use of efficient computer procedures, implementation of the expert system in the programming language R would be highly desirable. The aim of this paper is to present such an implementation, using an example

of the assignment of vegetation-plot records from the European Vegetation Archive (EVA) to the EUNIS habitat types. For testing the code, we provide a data set from the Tüxen Archive (Hoppe et al., 2005; Goral et al., 2018).

2 | METHODS

The algorithm was written in R (Version 3.6.3, R Foundation for Statistical Computing, Vienna, Austria). The R code is available and will be updated at https://git.loe.auf.uni-rostock.de/misc/ESy. Inputs are (i) the expert system file containing the logical definitions of vegetation types written in the ESy formal language and (ii) vegetation-plot data. The expert system file structure as it has evolved since the early 2000s (Kočí et al., 2003) has three sections: (1) species aggregation, (2) species groups and (3) type definitions (Tichý et al., 2019). Section 1 is a flexible solution to consider different taxon concepts (Jansen & Dengler, 2010) by combining different names, taxon levels or spelling to achieve operational taxa that can be used in the vegetation type definitions. The labels of these aggregated taxa are then used in Section 2 to define the species groups, which are the basic components for the vegetation type definitions given in Section 3. This section comprises a set of membership formulas that describe the assignment rules for plot records to be classified to individual vegetation types.

For the R implementation, the vegetation-plot data set has to consist of a data frame of plot identifiers ("RELEVE_NR"), species ("TaxonName"), and per cent cover ("Cover_Perc") in a narrow format (long table, stacked), similarly as used in the vegetation database management program TURBOVEG 2 (Hennekens & Schaminée, 2001). An additional header data file linked to the vegetation-plot data set by plot ID is required if the specific expert system makes use of site information, e.g. the plot location's longitude and latitude, altitude, as well as the assignment to country, ecoregion or other features (for the list in the case of the EUNIS classification see Chytrý et al., 2020b). For an example how to run the code we provide a test data set from the Tüxen Archive (Hoppe et al., 2005; Goral et al., 2018) comprising 10,717 plots from a wide range of vegetation types (see Electronic Appendix S1and the Readme.md file in the GIT repository). The data can be downloaded from https://www.vegetweb.de/#!quellendetails//1933). Species names were taxonomically harmonized with the taxval function from the *vegdata* package (Jansen & Dengler, 2010). This function performs a taxonomic valuation of species names according to taxonomic levels and synonyms, based on the reference list for Germany (germansl.infinitenature.org). Additionally, some header data for the EUNIS classification have been added. Results for single plots can be evaluated through helper functions eva and eva.type to show results for individual conditions and matching to type definitions. These helper functions correspond to the features in JUICE that allow an iterative optimization process for the development and iterative adjustment of formulas to minimize misclassifications (see Chytrý et al., 2020b).

The algorithm uses the following ten steps to assign plots in the data table of vegetation records (DT) to a vegetation type (Figure 1).

2.1 | Parse membership formulas

The R code transforms the membership formulas of Section 3 of the expert system file (Tichý et al., 2019: their Appendix S1) in a way that allows their evaluation as logical expressions in R. To this end, nested components are extracted, which in the following are evaluated simultaneously vector-wise for all plots, and finally combined to re-build the initial membership formula. In the example of the EUNIS membership formula below (R52 Forest fringe of acidic nutrient-poor soils), the membership expressions are limited by angle brackets (<…>) with the membership conditions within these membership expressions underlined:

((<##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils> AND <#02 +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils>) AND (<#TC +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils GR 25> OR <#SC +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils GE #$$>)) NOT <#TC Trees|#TC Shrubs GR 15>

2.2 | Complement membership expressions

In contrast to JUICE and TURBOVEG 3, where the membership formulas are applied sequentially, plot by plot, the R code solves all membership conditions independently from the context within membership expressions. This is particularly advantageous if the same membership conditions occur in different membership formulas, which is mostly the case in comprehensive expert system files, such as EUNIS-ESy (Chytrý et al., 2020b). However, this requires complementing the membership expressions if a logical operator and "the groups to be compared with" (the right-hand side of the expression) are missing. In the example above, this is the case for the ##Q expression, which defines that the sum of the square-root covers of a species group (##Q) should be greater than that of any other species group in the same set (the set "+04" in this case). These expressions only have a left-hand side condition, similarly to the expressions specifying the number of species (##D) or total cover (##C). The logical operator "GR NON" and a right-hand side condition is therefore added to the membership expression. The inserted logical operators and the inserted right-hand sides are shown in italics in the example of the membership formula of vegetation type R52:

((<##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils *GR NON ##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils*> AND <#02 +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils>) AND (<#TC +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils GR 25> OR <#SC +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils *GE #$$ EXCEPT +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils*>)) NOT <#TC Trees|#TC Shrubs GR 15>

In addition, if #$$ (the highest cover of any species not included in the species group) occurs on the right-hand side, the membership expression is complemented with EXCEPT and the group name on the right-hand side (see example above). Similarly, #T$ (total cover of all species in the record except those in the

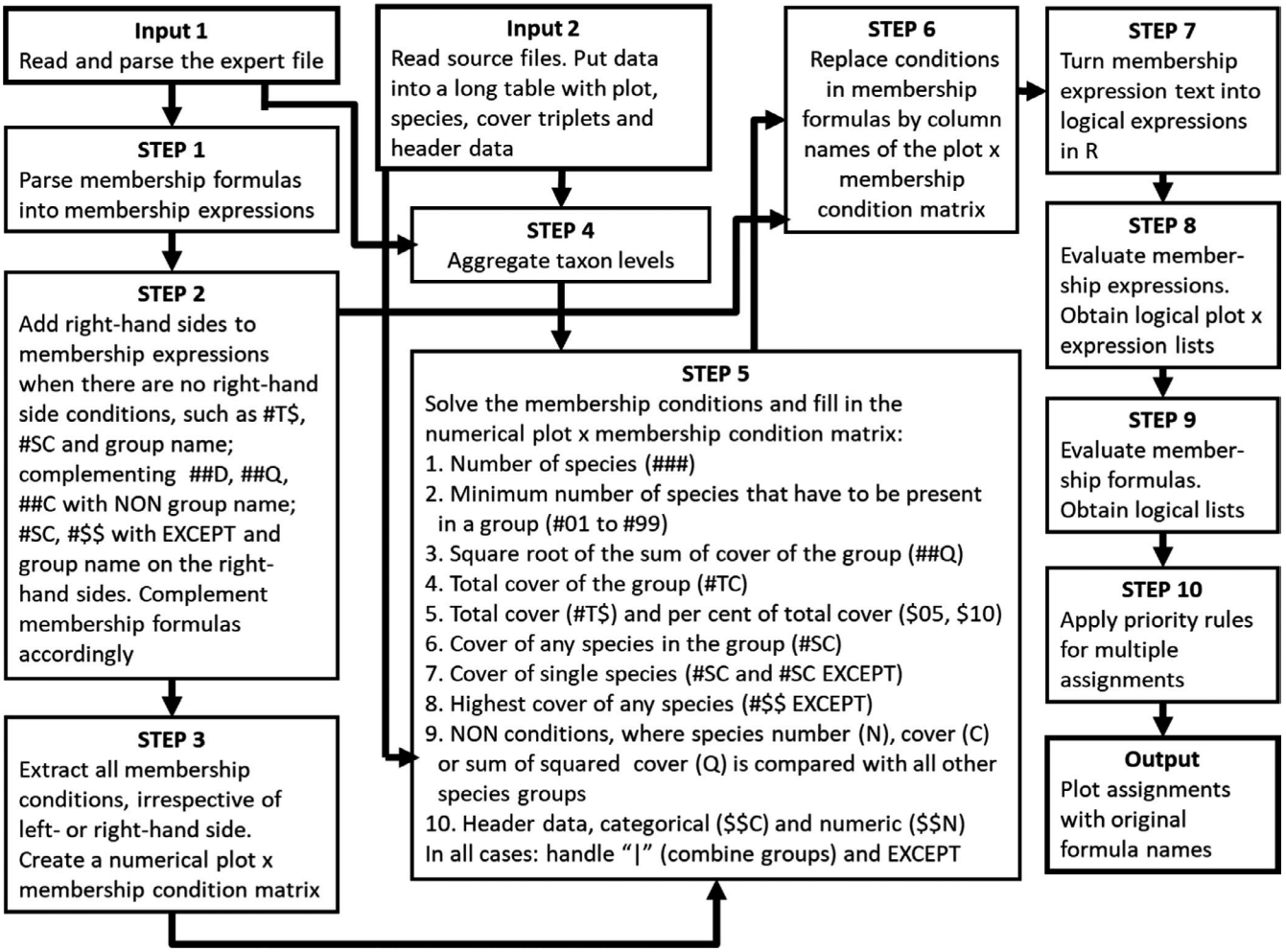

FIGURE 1 Implementation flow chart of the formal language of the expert system for vegetation classification in the programming language R

group mentioned) on the left-hand side of the membership expression has to be complemented with the group's name on the right-hand side. As a result, all membership expressions have a left- and right-hand side. Exceptions in the current implementation are the membership expressions ### and #01 to #99, which indicate the minimum number of species that have to be present from that group. As the #01 to #99 groups contain the operator (greater or equal) explicitly and ### groups implicitly (greater than or equal to half of the species of the group) and the right-hand condition is already in the group's name, they can be solved as a logical expression without the right-hand side. The additions to the membership expressions are then also implemented in the membership formulas.

2.3 | Parse membership expressions

All membership conditions are extracted from membership expressions, except for numeric conditions, which already comply to the required form of logical expressions in R. Using all the extracted membership conditions, an empty numeric plot × membership condition matrix is created.

2.4 | Aggregate taxa

Taxon names in data table of vegetation records (DT) are replaced by the names of the aggregated taxa under which they are listed in Section 1 of the expert system file (taxonomy.R).

2.5 | Solve the membership conditions

The membership conditions are then applied to the vegetation data in the data table (DT), and the numeric results are collected in the plot × membership condition matrix. Examples of the different types of membership conditions are shown in Figure 1, Step 5. Exceptions are membership conditions based on categorical header data ($$C), which require extracting the possible categories from the respective header field. Instead of storing the categories, their factor level numbers are

inserted in the plot × membership condition matrix. Other special cases are groups that have to be combined ("|", see the example above) or groups that are excluded (EXCEPT). For those cases, the condition has to be parsed, and the species of these groups are combined or excluded from the preceding part of the condition, respectively.

2.6 | Replace conditions in membership formulas

As all the membership conditions are now available as numeric values, these values can be inserted in the original formulas. In R, it is possible to do this vectorized, that is for all plots simultaneously, by referring to the column names in the plot × membership condition matrix. In our example, the membership condition "##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils" is stored for all plots in a column (which in the EUNIS-ESy is column "col531", with column number depending on the total number of membership conditions that are evaluated).

2.7 | Turn membership expression texts into logical expressions in R

The notations used in the expert system language are transformed into R logical operators, replacing "GR", "GE" and "EQ" with ">", ">=" and "==", respectively. As a result, the membership formula from the example above looks as follows:

((<col531 > col358> AND <col47>) AND (<col706 > 25> OR <col582 >= col175>)) NOT <col912 > 15>

2.8 | Evaluate membership expressions

The evaluation of logical expressions in R is also done vector-wise, solving all logical membership expressions in a single step, using the "eval" command in the lapply or mclapply function, and storing the results in a logical matrix. For example, "col288" holds the logical result (TRUE or FALSE) of "<##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils GR NON ##Q +04 R52-Forest-fringe-of-acidic-nutrient-poor-soils>" from the example above.

2.9 | Evaluate membership formulas

The logical membership formulas are solved in the same way in a single step, which is possible because the columns of the logical matrix only hold TRUE or FALSE entries:

((col288 & col289) & (col290 | col291)) & !col3).

The result is a list of the assignments to each membership formula (TRUE or FALSE) for every plot.

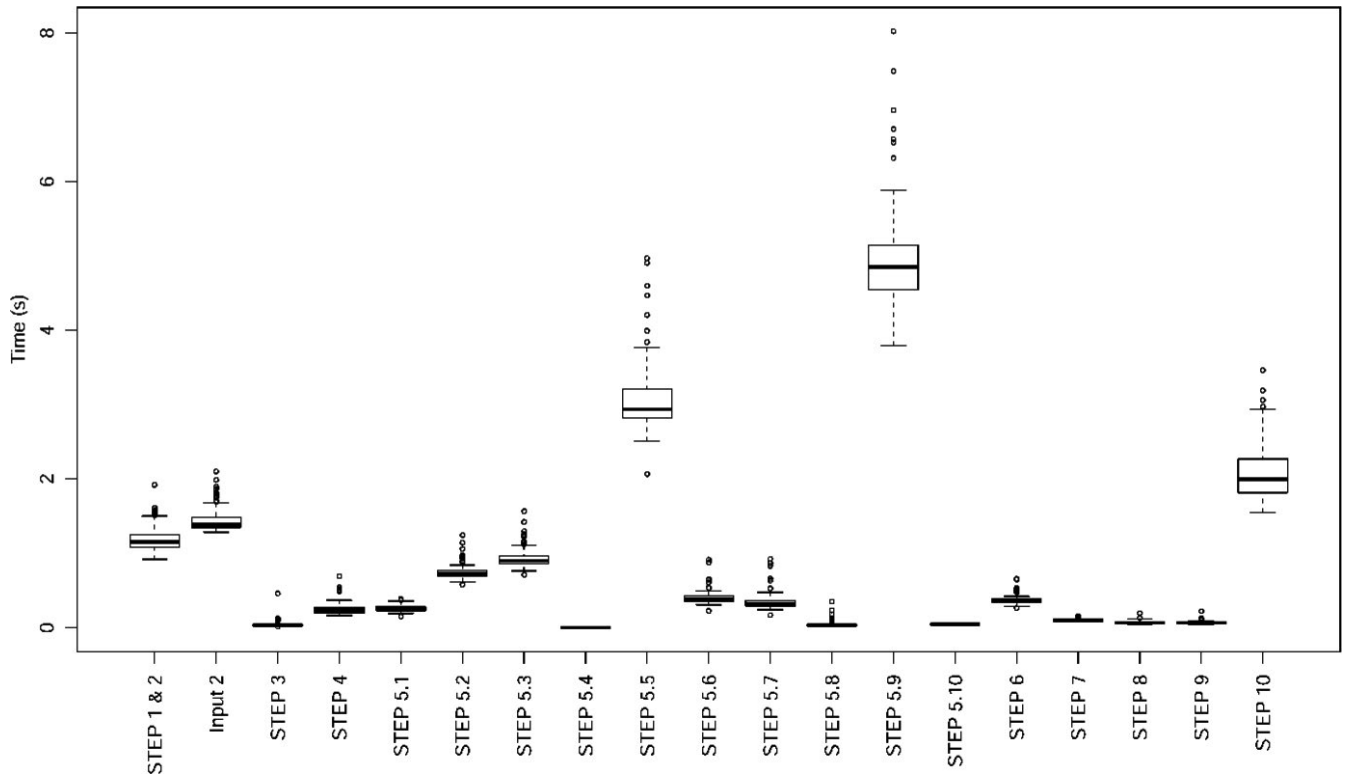

FIGURE 2 Average run time of 100 randomized subsets of 10,000 plots for different steps of the classification process. An increase of the number of plots leads to a linear increase in steps 5–10, whereas steps 1–4 depend only on the size of the expert system file. The average total run time across these 100 runs is 16.5 s (using three cores)

2.10 | Apply priority rules

As a plot may be assigned to more than one vegetation type, priority rules are applied (Chytrý et al., 2020b). The priority level of the vegetation type can be parsed from the input line with the type's name from the expert system file. In the end, the plot is assigned to the vegetation type with the highest priority of all that were found to be TRUE in step 9. The plots with multiple assignments at the same hierarchical level are indicated with "+", plots without valid assignment to any vegetation type with "?".

3 | RESULTS AND DISCUSSION

We tested the R implementation of the expert system on a data set of 1.4 Mio. European vegetation plots from the EVA database and the expert system EUNIS-ESy developed for the EUNIS Habitat Classification (Chytrý et al., 2020b). Figure 2 shows the run time of the classification process for 100 randomized subsets of 10,000 plots each. The most time-demanding steps are 5.5 and 5.9. In step 5.5, the #T$ conditions are evaluated, which is the total cover of all species in the plot, excluding the species of the group involved in the comparison. This step is time-demanding mainly because of possible EXCEPT conditions, which have to be considered here. In step 5.9, the #TC conditions are evaluated, which requires calculating the total cover of all other species groups except those of the target group. However, there is undoubtedly potential to speed up the current code. With the simultaneous vector-based processing in R, the time demand for the ESy classification scales more or less linearly with the number of records classified. Thus, the R version now also allows iterative cycles of classifying future global data sets of several million plot records.

The assignment for the worked-out example of vegetation types from Germany is shown in Figure 3. Out of the 10,717 vegetation-plot records, 10,084 (94%) were unambiguously assigned to a single EUNIS habitat type, 7,810 of them to level 3 habitat types and 2,274 to habitat groups (level 1). In contrast, 619 records (5.8% of all 10,717 records), did not fulfill the criteria of any of the EUNIS habitat types, which is only slightly higher than the amount of un-assignable plot records in the European EUNIS-ESy classification (4.7% of 1,261,373

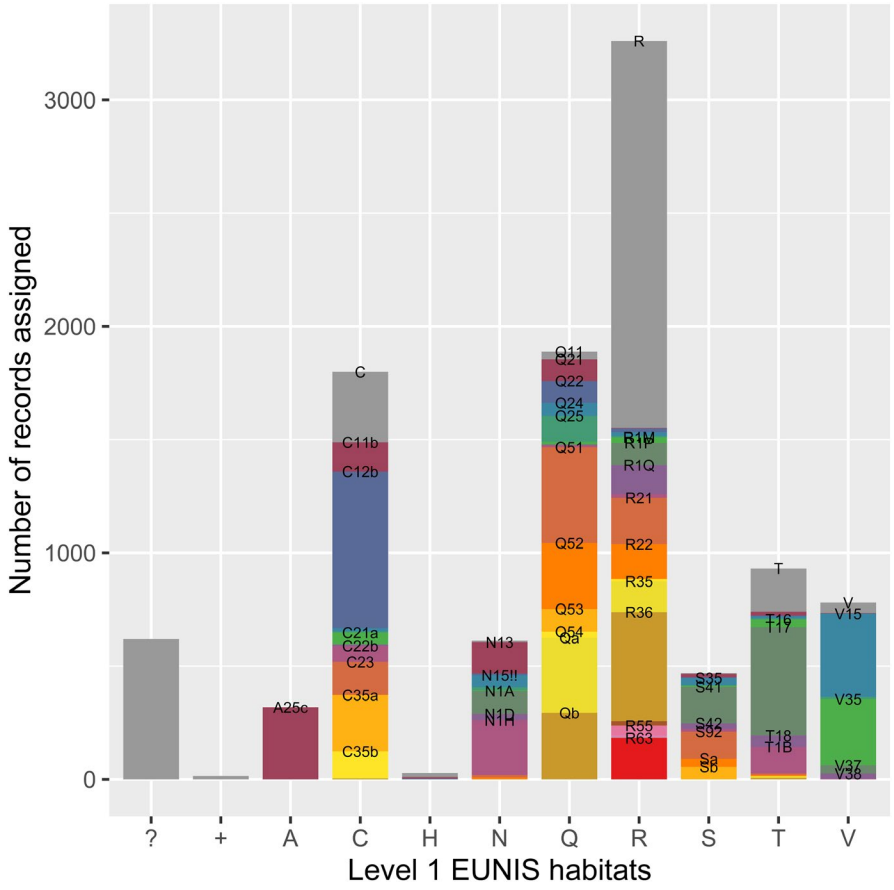

FIGURE 3 Results of the classification by the EUNIS-ESy expert system applied to a test data set from the Tüxen Archive (project "100716 Hoppe, 2005" from vegetweb.de). ?: plots not assigned to any level 3 EUNIS habitat; +: assigned to more than one level 3 EUNIS habitat; A, marine habitats; C, inland surface waters; H, inland unvegetated or sparsely vegetated habitats; N, coastal habitats; Q, wetlands; R, grasslands and lands dominated by forbs, mosses or lichens; S, heathlands, scrub and tundra; T, forests and other wooded land; V, vegetated man-made habitats. Labels for EUNIS habitats were only printed at the top of the corresponding bar section when the number of assigned records was larger than or equal to 20

plot records, Chytrý et al., 2020b). Furthermore, 14 records (0.13%) were assigned to more than one type, corresponding to 0.2% in Chytrý et al. (2020b).

The expert system language for vegetation classification (Tichý et al., 2019) has been used in numerous applications, including the complete national vegetation classification of the Czech Republic to the association level (Chytrý, 2007–2013; Chytrý et al., 2020a) and continental-scale vegetation surveys (e.g. Douda et al., 2016; Peterka et al., 2017; Willner et al., 2017; Marcenò et al., 2018; Landucci et al., 2020). It is increasingly used in European nature conservation, in particular through its application in the definitions of habitat types of EUNIS, a standard classification of European habitats used by the European Environment Agency (Chytrý et al., 2020b). Although the use of vegetation classification expert systems was mainly restricted to Europe so far, perhaps mainly due to the availability of large data from vegetation plots on this continent (Chytrý et al., 2016), there are also examples of their application elsewhere (e.g. Li et al., 2013). The recent accumulation of such data from other continents in the sPlot database (Bruelheide et al., 2019) opens various future applications options.

An advantage of the R implementation is that membership formulas are not only readable in R but can also be produced as machine-written membership formulas by R scripts, for example as an output of any classification algorithm in R. Our vision is that classification programs (such as Cocktail clustering, Bruelheide, 2016) should provide a machine-readable ESy file as standard output.

For applications of the expert system, it is of key importance that the user adjusts the taxon nomenclature in the input data file to match the nomenclature used in the expert system. Also, if the expert system uses information from the header data such as plot location, it is important that all the required variables are available in the input file.

Although there were earlier attempts of applying the expert system language as an R code (Li et al., 2013), the code presented here reflects the up-to-date state of the expert system language as described by Tichý et al. (2019) with possibilities of defining vegetation types at different hierarchical levels and the use of assignment rules based on the site information in addition to those based on species composition and cover. The availability of the classification tool in R, in addition to the previous implementations in JUICE (Tichý, 2002) and TURBOVEG 3 (Hennekens, 2015), enhances the flexibility of using the expert system classifications and allows combining it with other analytical tools for ecological and environmental research and nature conservation applications.

ACKNOWLEDGEMENTS

We thank Ute Jandt for her input in explaining assignment rules and careful proofreading, and Sebastian Schmidtlein, Attila Lengyel and two anonymous reviewers for their constructive comments. Open-access funding was enabled and organized by Project DEAL.

AUTHOR CONTRIBUTIONS

HB conceived the idea and developed the initial version of the R program, which FJ and HB then jointly improved. LT and MC provided insight into the ESy language and the EUNIS-ESy expert system. All authors tracked errors in the code and contributed to paper writing.

DATA AVAILABILITY STATEMENT

The R code is available at https://git.loe.auf.uni-rostock.de/misc/ESy.

ORCID

Helge Bruelheide https://orcid.org/0000-0003-3135-0356
Lubomír Tichý https://orcid.org/0000-0001-8400-7741
Milan Chytrý https://orcid.org/0000-0002-8122-3075
Florian Jansen https://orcid.org/0000-0002-0331-5185

REFERENCES

Bruelheide, H. (1997) Using formal logic to classify vegetation. *Folia Geobotanica et Phytotaxonomica*, 32, 41–46. https://doi.org/10.1007/BF02803883

Bruelheide, H. (2016) Cocktail clustering – a new hierarchical agglomerative algorithm for extracting species groups in vegetation databases. *Journal of Vegetation Science*, 27, 1297–1307. https://doi.org/10.1111/jvs.12454

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S.M., Chytrý, M. et al. (2019) sPlot – a new tool for global vegetation analyses. *Journal of Vegetation Science*, 30, 161–186. https://doi.org/10.1111/jvs.12710

Bruelheide, H. & Jandt, U. (1997) Demarcation of communities in large databases. *Phytocoenologia*, 27, 141–159.

Chytrý, M. (Ed.) (2007–2013) *Vegetation of the Czech Republic 1–4*. (in Czech). Praha: Academia. https://botzool.cz/vegsci/vegetationCR

Chytrý, M., Hennekens, S.M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F. et al. (2016) European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science*, 19, 173–180. https://doi.org/10.1111/avsc.12191

Chytrý, M., Tichý, L., Boublík, K., Černý, T., Douda, J., Hájek, M. et al. (2020a) CzechVeg-ESy: Expert system for automatic classification of vegetation plots from the Czech Republic. *Zenodo*, https://doi.org/10.5281/zenodo.3605562

Chytrý, M., Tichý, L., Hennekens, S.M., Knollová, I., Janssen, J.A.M., Rodwell, J.S. et al. (2020b) EUNIS Habitat Classification: expert system, characteristic species combinations and distribution maps of European habitats. *Applied Vegetation Science*, 23, 648–675. https://doi.org/10.1111/avsc.12519

De Cáceres, M. & Wiser, S.K. (2012) Towards consistency in vegetation classification. *Journal of Vegetation Science*, 23, 387–393. https://doi.org/10.1111/j.1654-1103.2011.01354.x

Douda, J., Boublík, K., Slezák, M., Biurrun, I., Nociar, J., Havrdová, A. et al. (2016) Vegetation classification and biogeography of European floodplain forests and alder carrs. *Applied Vegetation Science*, 19, 147–163. https://doi.org/10.1111/avsc.12201

Goral, F., Hoppe, A. & Bergmeier, E. (2018) Open Access zu 13 500 europäischen Vegetationsaufnahmen aus dem Reinhold-Tüxen-Archiv. *Tuexenia*, 38, 297–304.https://doi.org/10/gg8zrh

Hennekens, S.M. (2015) Turboveg 3 – A gateway to EVA and other databases. In: Chytrý, M., Zelený, D. and Hettenbergerová, E. (Eds.) *58th Annual Symposium of the International Association for Vegetation Science: Understanding broad-scale vegetation patterns – Abstracts*. Brno: Masaryk University, p. 152.

Hennekens, S. M., & Schaminée, J. H. J. (2001). TURBOVEG, a comprehensive data base management system for vegetation data. *Journal of Vegetation Science*, 12(4), 589–591. https://doi.org/10.2307/3237010

Hoppe, A. (2005) Das Reinhold-Tüxen-Archiv am Institut für Geobotanik der Universität Hannover. *Tuexenia*, 25, 463–474.

Jansen, F. & Dengler, J. (2010) Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science*, 21, 1179–1186. https://doi.org/10.1111/j.1654-1103.2010.01209.x

Kočí, M., Chytrý, M. & Tichý, L. (2003) Formalized reproduction of an expert-based phytosociological classification: a case study of subalpine tall-forb vegetation. *Journal of Vegetation Science*, 14, 601–610. https://doi.org/10.1111/j.1654-1103.2003.tb02187.x

Landucci, F., Šumberová, K., Tichý, L., Hennekens, S., Aunina, L., Biță-Nicolae, C. et al. (2020) Classification of the European marsh vegetation (Phragmito-Magnocaricetea) to the association level. *Applied Vegetation Science*, 23, 297–316. https://doi.org/10.1111/avsc.12484

Landucci, F., Tichý, L., Šumberová, K. & Chytrý, M. (2015) Formalized classification of species-poor vegetation: a proposal of a consistent protocol for aquatic vegetation. *Journal of Vegetation Science*, 26, 791–803. https://doi.org/10.1111/jvs.12277

Li, C.-F., Chytrý, M., Zelený, D., Chen, M.-Y., Chen, T.-Y., Chiou, C.-R. et al. (2013) Classification of Taiwan forest vegetation. *Applied Vegetation Science*, 16, 698–719. https://doi.org/10.1111/avsc.12025

Marcenò, C., Guarino, R., Loidi, J., Herrera, M., Isermann, M., Knollová, I. et al. (2018) Classification of European and Mediterranean coastal dune vegetation. *Applied Vegetation Science*, 21, 533–559. https://doi.org/10.1111/avsc.12379

Peterka, T., Hájek, M., Jiroušek, M., Jiménez-Alfaro, B., Aunina, L., Bergamini, A. et al. (2017) Formalized classification of European fen vegetation at the alliance level. *Applied Vegetation Science*, 20, 124–142. https://doi.org/10.1111/avsc.12271

Tichý, L. (2002) JUICE, software for vegetation classification. *Journal of Vegetation Science*, 13, 451–453. https://doi.org/10.1111/j.1654-1103.2002.tb02069.x

Tichý, L., Chytrý, M. & Landucci, F. (2019) GRIMP: A machine-learning method for improving groups of discriminating species in expert systems for vegetation classification. *Journal of Vegetation Science*, 30, 5–17. https://doi.org/10.1111/jvs.12696

Willner, W., Jiménez-Alfaro, B., Agrillo, E., Biurrun, I., Campos, J.A., Čarni, A. et al. (2017) Classification of European beech forests: a Gordian Knot? *Applied Vegetation Science*, 20, 494–512. https://doi.org/10.1111/avsc.12299

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

Appendix S1. R code of the expert system (ESy-release_v1.0) for the assignment of vegetation plots to vegetation types and test data (from the Tüxen Archive, Hoppe et al., 2005; Gotal et al., 2018).

How to cite this article: Bruelheide H, Tichý L, Chytrý M, Jansen F. Implementing the formal language of the vegetation classification expert systems (ESy) in the statistical computing environment R. *Appl Veg Sci*. 2021;24:e12562. https://doi.org/10.1111/avsc.12562