REPORT

Probabilistic key for identifying vegetation types in the field: A new method and Android application

Lubomír Tichý (iD) | Milan Chytrý (iD)

Department of Botany and Zoology, Masaryk University, Brno, Czech Republic

Correspondence
Lubomír Tichý, Department of Botany and Zoology, Masaryk University, Brno, Czech Republic.
Email: tichy@sci.muni.cz

Abstract

Quick identification of vegetation types in the field, based on species composition but not requiring time-consuming plot sampling, is often needed for vegetation mapping, conservation assessment, teaching and other applications of vegetation classification. Here, we propose a new method that identifies the probability of belonging to the units of an established vegetation classification for vegetation stands encountered in the field. The method is based on calculating the probability that a few species observed in the field would co-occur in a priori defined vegetation types, using the existing information on species occurrence frequency in these types. The method has been implemented in a freely available Android application called Probabilistic Vegetation Key, which makes it possible to employ it in the field using smartphones or tablets, even in the absence of internet access.

KEYWORDS
Android, field survey, identification, smartphone, software, species composition, tablet, vegetation classification, vegetation survey, vegetation type

1 | INTRODUCTION

Vegetation science has a long tradition of classifying vegetation into types defined by species composition (De Cáceres et al., 2015; Mucina et al., 2016). Such vegetation types can be used for vegetation mapping, monitoring, conservation planning and assessment, or for defining study systems in basic research. The key question for the users of vegetation classification systems is how to identify vegetation types, either in the field or in databases of vegetation plots. Several methods have been developed for the identification of vegetation types in databases, some of them based on similarity in species composition between a single plot record and a set of plots previously classified to the types (Gégout & Coudun, 2012; Hill, 1989; Tichý, 2005; van Tongeren, Gremmen, & Hennekens, 2008), others based on expert systems comprising formal definitions of types (Bruelheide, 1997; Kočí, Chytrý, & Tichý, 2003; Landucci, Tichý, Šumberová, & Chytrý, 2015; Tichý, Chytrý, &

Landucci, 2019). However, there are hardly any tools that would enable quick identification of vegetation types directly in the field. Vegetation scientists have long recognized that usually only relatively small subsets of species from the total species composition are important for characterizing or discriminating vegetation types (Dale, Beatrice, Venanzoni, & Ferrari, 1986; Jancey, 1979). Therefore, a few species should be sufficient to identify vegetation types in the field, provided they are characteristic of the considered vegetation types. Identification based on a few species found on the spot would be very useful for vegetation or habitat mapping, conservation assessment and other tasks that do not allow enough time for sampling complete species composition and abundances in vegetation plots.

Here, we introduce a new method to identify vegetation types based on a few observed species, called Probabilistic Vegetation Key, and provide a software application for smartphones or tablets that allows the use of this method in the field.

2 | IDENTIFICATION METHOD

The identification of vegetation types in the Probabilistic Vegetation Key is based on the probability that an observed set of species co-occurs in a particular vegetation type. Existing systems of vegetation classification are usually characterized by synoptic tables of species composition, which include occurrence frequencies (also called constancies) of each species in each vegetation type. These frequencies can be used as estimations of the probability of occurrence of each individual species in each type. For more than one species, the probability of their co-occurrence in a stand of a vegetation type can be defined as a product of occurrence probabilities (relative occurrence frequencies) of individual species:

$$V_i = p_1 \cdot p_2 \cdot p_3 \dots p_n$$

where V_i is the probability of co-occurrence of n selected species in vegetation type i and p_1, p_2, p_3, ... p_n are the probabilities of occurrence of species 1, 2, 3, ... n in this type, expressed as their relative occurrence frequencies in this type on the scale from 0 to 1.

Using this approach, a vegetation type can be erroneously excluded from consideration if a user selects a generalist species that can occur in the type but is not present in the list of species for this type, or if a user enters a misidentified species not present in the type. To make the algorithm more robust to such errors, the minimum relative frequency for all species was arbitrarily set to 0.0001, which does not exclude any vegetation type because of a single species.

The probabilities of co-occurrence of the selected set of species in vegetation types 1, 2, 3... k (V_1, ... V_k) are summed across all k vegetation types defined in the classification. The relative probability R_i (in percentages) estimated for the vegetation type i is then calculated as the species co-occurrence probability V_i for this type divided by the sum:

$$R_i(\%) = 100 \cdot V_i / \sum_{j=1}^{k} V_j$$

The vegetation type for which this value is the highest is the most plausible identification of the vegetation stand in which the observed set of species was found. Unlike probability V, which decreases with an increasing number of observed species for all vegetation types, the relative probabilities across all vegetation types R_1, ... R_k always add up to 100%.

The probabilistic identification is based on species composition. In addition, physiognomically different vegetation types, such as forest vs non-forest, can be analyzed independently, which decreases the misidentification rate. The classification accuracy could be further improved by adding environmental or geographical information on individual types. However, as each of such variables has different importance for discriminating different vegetation types, we decided not to include them into the identification procedures. Instead, they can be used as external criteria for verification of the classification results.

3 | A TEST WITH REAL DATA

We used data from the national vegetation classification of the Czech Republic (Chytrý, 2007–2013), which includes 496 phytosociological associations. Occurrence frequencies of all species in all associations were calculated based on vegetation plots from the Czech National Phytosociological Database (Chytrý & Rafajová, 2003), which were classified to associations by the expert system developed as a part of the national vegetation classification, and then resampled within geographic strata nested within associations in order to reduce local oversampling. The resampled data set contained 30,115 plots.

To test the degree of success of the identification of associations, we prepared two selections of species. First, we randomly selected 10,000 vegetation plots (separately for forest and non-forest plots), and in each of them, we randomly selected groups of 2, 3, 4, 5 and 6 co-occurring species. These species combinations included both rare and common species, both specialists and generalists, and species typical of different associations. We also recorded the assignment of each plot to the association. This data set represented the case when a person records species in the field, having no idea about their relationship to vegetation types. Second, we asked fourteen vegetation scientists to select three phytosociological associations familiar to them and list six species for each association in decreasing order of their estimated importance for differentiation of the association. This data set represented the case when a person has field experience with the studied vegetation, can recognize local vegetation types and understands which species characterize each type, but does not know the name of the type according to the regional/national vegetation classification system. For both data sets, we calculated the mean relative probability of the selected species combination to co-occur in the respective association, and ordered the associations by this probability (Table 1). An example of a specific species combination and related probabilities is shown in Appendix S1.

For both species selections, and for both forest and non-forest vegetation, the probability of correct assignment to the association increased with the number of species used. With the expert-based species selections, the method was able to identify the correct association out of 496 (or at least to place it on the second position) using only 2–4 species. When species were selected randomly, more species were usually needed for correct identification, but this test made it clear that even a botanist with no idea about the classification of vegetation types is able to restrict a large number of vegetation types to a small group of possibly correct types. The final identification can be made by using additional information (e.g., physiognomy, ecology, distribution, and a full list of indicator species) about the vegetation types that were identified as the most probable.

TABLE 1 The success of the identification of phytosociological associations using lists of 2, 3, … 6 species, either selected randomly from a vegetation-plot database or selected by vegetation scientists with a priori knowledge of these associations; the mean order of the correct association on the list of 496 associations is based on the decreasing mean relative probability that the selected group of species co-occurs in this association

Randomly selected species from the vegetation database					
Forest vegetation			**Non-forest vegetation**		
Number of species	Mean relative probability (%)	Mean order of the correct association	Number of species	Mean relative probability (%)	Mean order of the correct association
2	12.1	6.2	2	11.8	10.4
3	20.8	4.6	3	21.9	6.8
4	32.0	3.1	4	24.7	4.2
5	41.4	2.9	5	33.2	3.4
6	47.9	2.0	6	39.2	3.5
Expert-based selection of species (14 experts)					
Forest vegetation			**Non-forest vegetation**		
Number of species	Mean relative probability (%)	Mean order of the correct association	Number of species	Mean relative probability (%)	Mean order of the correct association
2	34.0	1.8	2.0	64.3	1.1
3	46.9	1.6	3.0	77.9	1.0
4	53.6	1.5	4.0	85.5	1.0
5	57.7	1.5	5.0	87.4	1.0
6	69.6	1.5	6.0	88.7	1.1

4 | APPLICATION

The Probabilistic Vegetation Key can identify vegetation types or habitat types at any level of the classification hierarchy. It does not require recording full species composition in vegetation plots. Therefore, it is useful for fast identification of vegetation types directly in the field, for example in vegetation mapping, for a preliminary exploration of vegetation diversity in an area, and in field courses of vegetation science. To enable the application of the method in the field, an Android application that can be used in mobile phones or tablets was developed by the first author of this report. This application, called the Probabilistic Vegetation Key, is freely available at https://play.google.com/store/apps/details?xml:id=com.test.tichy.vegkey. Its structure and interconnection are described in Appendix S2. This application uses several predefined databases. One of them is a list of species, from which the user can select those observed in the field, specify which of them have a high cover and whether they were found in forest or non-forest vegetation. Based on this information, the program returns a sorted list of vegetation types with their probabilities, with the most probable types on top. All vegetation types are additionally described by diagnostic, constant and dominant species together with brief characteristics of their physiognomy and ecology. The application does not require an internet connection, which is helpful mainly in the field where internet access may not be available. At present, the program contains data for the national vegetation classification of the Czech Republic at the level of associations, and consequently, it is only applicable for this classification system. However, any other system of vegetation or habitat classification can be added, provided a list of species with occurrence frequency (and optionally also a list of species with high cover) is provided for each vegetation or habitat type. To enable the use of this application in other countries and regions, the source code for Android Studio has been made available in the Figshare repository (Tichý, 2019).

ORCID

Lubomír Tichý https://orcid.org/0000-0001-8400-7741
Milan Chytrý https://orcid.org/0000-0002-8122-3075

REFERENCES

Bruelheide, H. (1997). Using formal logic to classify vegetation. *Folia Geobotanica et Phytotaxonomica*, *32*(1), 41–46. https://doi.org/10.2307/4181516

Chytrý, M. (Ed.) (2007–2013). *Vegetace České republiky 1–4*. Praha, Czech Republic: Academia.

Chytrý, M., & Rafajová, M. (2003). Czech National Phytosociological Database: basic statistics of the available vegetation-plot data. *Preslia*, *75*, 1–15.

Dale, M. B., Beatrice, M., Venanzoni, R., & Ferrari, C. (1986). A comparison of some methods of selecting species in vegetation analysis. *Coenoses*, *1*, 35–51.

De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., … Wiser, S. K. (2015). A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science*, *18*, 543–560. https://doi.org/10.1111/avsc.12179

Gégout, J.-C., & Coudun, C. (2012). The right relevé in the right vegetation unit: a new typicality index to reproduce expert judgement with an automatic classification programme. *Journal of Vegetation Science*, *23*, 24–32. https://doi.org/10.1111/j.1654-1103.2011.01337.x

Hill, M. O. (1989). Computerized matching of relevés and association tables, with an application to the British National Vegetation Classification. *Vegetatio*, *83*, 187–194. https://doi.org/10.1007/bf00031691

Jancey, R. C. (1979). Species ordering on a variance criterion. *Vegetatio*, *39*, 59–63. https://doi.org/10.1007/bf00055329

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., 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

Mucina, L., Bültmann, H., Dierßen, K., Theurillat, J.-P., Raus, T., Čarni, A., … Tichý, L. (2016). Vegetation of Europe: Hierarchical floristic classification system of vascular plant, lichen, and algal communities. *Applied Vegetation Science*, *19*(Suppl.), 3–264. https://doi.org/10.1111/avsc.12257

Tichý, L. (2005). New similarity indices for the assignment of relevés to the vegetation units of an existing phytosociological classification. *Plant Ecology*, *179*, 67–72. https://doi.org/10.1007/s11258-004-5798-8

Tichý, L. (2019). Source code for probabilistic vegetation key. Retrieved from Figshare. https://doi.org/10.6084/m9.figshare.8256164

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

van Tongeren, O., Gremmen, N., & Hennekens, S. (2008). Assignment of relevés to pre-defined classes by supervised clustering of plant communities using a new composite index. *Journal of Vegetation Science*, *19*, 525–536. https://doi.org/10.3170/2008-8-18402

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1. An example of the identification process of an association with successive addition of one to six species.

Appendix S2. Data flow in the Android application "Probabilistic Vegetation Key".

How to cite this article: Tichý L, Chytrý M. Probabilistic key for identifying vegetation types in the field: A new method and Android application. *J Veg Sci*. 2019;30:1035–1038. https://doi.org/10.1111/jvs.12799

Supporting information to the paper

Tichý, L., & Chytrý, M. (2019). Probabilistic key for identifying vegetation types in the field: a new method and Android application. *Applied Vegetation Science*.

Appendix S1 An example of the identification process of an association with successive addition of one to six species. The correct association is *Sorbo torminalis-Quercetum*. Percentage probabilities for forest associations from the Czech national vegetation classification, sorted in decreasing order, are shown, excluding those with a probability lower than 3%. After adding the first species, the correct association scored second in the order of probabilities, but the probability of correct assignment was very low (6.0%). Since the addition of the third species, the correct association scored first, and after adding the sixth species, the probability of correct assignment increased to 59.4%.

Quercus petraea agg.		&	Euphorbia cyparissias		&	Vincetoxicum hirundinaria		&	Genista tinctoria		&	Anthericum ramosum		&	Festuca ovina	
6.2%	Genisto pilosae-Quercetum petraeae	16.2%	Genisto pilosae-Quercetum petraeae	18.3%	Sorbo torminalis-Quercetum	31.2%	Sorbo torminalis-Quercetum	38.3%	Sorbo torminalis-Quercetum	59.4%	Sorbo torminalis-Quercetum
6.0%	**Sorbo torminalis-Quercetum**	11.2%	**Sorbo torminalis-Quercetum**	17.5%	Lithospermo purpureocaerulei-Quercetum pubescentis	26.4%	Lithospermo purpureocaerulei-Quercetum pubescentis	28.5%	Lithospermo purpureocaerulei-Quercetum pubescentis	17.5%	Genisto pilosae-Quercetum petraeae
5.9%	Primulo veris-Carpinetum betuli	8.3%	Lithospermo purpureocaerulei-Quercetum pubescentis	16.9%	Genisto pilosae-Quercetum petraeae	9.3%	Genisto pilosae-Quercetum petraeae	9.5%	Seslerio albicantis-Tilietum cordatae	8.3%	Seslerio albicantis-Tilietum cordatae
5.8%	Viscario vulgaris-Quercetum petraeae	8.2%	Euphorbio-Quercetum	8.5%	Seslerio albicantis-Tilietum cordatae	9.2%	Viscario vulgaris-Quercetum petraeae	8.0%	Genisto pilosae-Quercetum petraeae	6.9%	Festuco-Pinetum sylvestris
5.5%	Quercetum pubescenti-roboris	6.3%	Lathyro collini-Quercetum pubescentis	7.2%	Festuco-Pinetum sylvestris	6.9%	Seslerio albicantis-Tilietum cordatae	5.8%	Festuco-Pinetum sylvestris	3.6%	Viscario vulgaris-Quercetum petraeae
5.4%	Melico pictae-Quercetum roboris	6.3%	Festuco-Pinetum sylvestris	5.8%	Quercetum pubescenti-roboris	5.3%	Festuco-Pinetum sylvestris	4.5%	Lathyro collini-Quercetum pubescentis	3.0%	Lathyro collini-Quercetum pubescentis
5.3%	Luzulo luzuloidis-Quercetum petraeae	6.1%	Seslerio albicantis-Tilietum cordatae	5.8%	Euphorbio-Quercetum	3.8%	Lathyro collini-Quercetum pubescentis				
5.3%	Vaccinio vitis-idaeae-Quercetum roboris	5.5%	Primulo veris-Carpinetum betuli	4.1%	Lathyro collini-Quercetum pubescentis						
5.2%	Euphorbio-Quercetum	5.3%	Melico pictae-Quercetum roboris								
4.5%	Galio sylvatici-Carpinetum betuli	5.0%	Quercetum pubescenti-roboris								
3.8%	Carici pilosae-Carpinetum betuli	4.1%	Viscario vulgaris-Quercetum petraeae								
3.7%	Seslerio albicantis-Tilietum cordatae	3.1%	Pruno spinosae-Ligustretum vulgaris								

Supporting information to the paper

Tichý, L., & Chytrý, M. (2019). Probabilistic key for identifying vegetation types in the field: a new method and Android application. *Applied Vegetation Science*.

Appendix S2. Data flow in the Android application ‚Probabilistic Vegetation Key'.

Selected species
Data manually selected by the user:

Quercus petraea agg.
Euphorbia cyparissias
Vincetoxicum hirundinaria

Species list
Species code (3+3 chars), Spec. name, Spec. ID

```
ACEPLA  Acer platanoides          5
ACEPSE  Acer pseudoplatanus       6
ACHASP  Achillea asplenifolia     7
ACHMIL  Achillea millefolium agg. 8
ACHNOB  Achillea nobilis          9
ACHPAN  Achillea pannonica        10
ACHPTA  Achillea ptarmica         11
ACHSET  Achillea setacea          12
```
file: species.txt

List of all associations (alliances)

Published vegetation association code, full vegetation association name, (line = Vegetation association ID)

```
ACA01;;;Saxifrago oppositifoliae-Festucetum versicoloris
ACA02;;;Saxifrago paniculatae-Agrostietum alpinae
ADA01;;;Sphagno compacti-Molinietum caeruleae
ADA02;;;Crepido conyzifoliae-Calamagrostietum villosae
ADA03;;;Violo sudeticae-Deschampsietum cespitosae
ADB01;;;Bupleuro longifoliae-Calamagrostietum arundinaceae
ADC01;;;Salici silesiacae-Betuletum carpaticae
```
file: syntaxa.txt

Synoptic table of the complete Czech National Vegetation Classification

Species ID, association ID, No. of plots assigned to the association, No. of plots with the species, No. of plots with the species occurring as a dominant (cover > 25%)

```
148,476,9,1,0
148,477,232,81,10
148,478,109,33,2
148,479,145,56,1
148,480,50,18,3
148,481,41,25,0
148,482,18,2,0
```
file: ttbl_25.txt

Android application:
https://play.google.com/store/apps/details?id=com.test.tichy.vegkey&hl=en

Free code for Android Studio and data files:
https://doi.org/10.6084/m9.figshare.8256164.v2

List of the most probable vegetation types

Algorithm result:

18.3% *Sorbo torminalis-Quercetum*
17.5% *Lithospermo purpureocaerulei-Quercetum pubescentis*
16.9% *Genisto pilosae-Quercetum petraeae*
8.5% *Seslerio albicantis-Tilietum cordatae*
7.2% *Festuco-Pinetum sylvestris*
5.8% *Quercetum pubescenti-roboris*
5.8% *Euphorbio-Quercetum*
4.1% *Lathyro collini-Quercetum pubescentis*

Diagnostic, constant and dominant species of the selected vegetation type and its basic characteristics

Unit Description
LCC01 Sorbo torminalis-Quercetum

Possible Habitat:
(Habitat Catalogue of the Czech Republic)
L6.5B Acidophilous thermophilous oak forests

Diagnostic Species:
Anthericum ramosum, Campanula persicifolia, Digitalis grandiflora, Festuca ovina, Genista tinctoria, Hieracium lachenalii, Hieracium murorum, Hieracium sabaudum s. l., Hylotelephium maximum, Poa nemoralis, Polygonatum odoratum, Quercus petraea agg., Silene nutans, Tanacetum corymbosum, Trifolium alpestre, Vincetoxicum hirundinaria, Viscaria vulgaris

Constant Species:
Carpinus betulus, Quercus petraea agg., Anthericum ramosum, Campanula persicifolia, Euphorbia cyparissias, Festuca ovina, Genista tinctoria, Hieracium lachenalii, Hieracium murorum, Hieracium sabaudum s. l., Hylotelephium maximum, Hypericum perforatum, Luzula luzuloides, Poa nemoralis, Polygonatum odoratum, Silene nutans, Tanacetum corymbosum, Veronica vindobonensis, Vincetoxicum hirundinaria, Viscaria vulgaris, Hypnum cupressiforme s. l.

Dominant Species:
Quercus petraea agg, Brachypodium pinnatum, Carex humilis, Festuca ovina, Poa nemoralis

Characteristics:
This association is usually dominated by Quercus petraea agg. The shrub layer is sparse and the herb layer is composed of acidotolerant, light-demanding thermophilous species, in some stands also with an admixture of species of mesic forests. A moss layer is often well developed. This vegetation occurs on south-facing slopes of poorly-weathering acidic rocks such as granite, gneiss, shale or volcanic rocks. Soils are usually shallow rankers or cambisols. In the Czech Republic this vegetation occurs mainly on volcanic hills of northern Bohemia and in deep river valleys of central and southern Bohemia and western Moravia.

files: dgdr.txt, codr.txt, dmdr.txt, assoc_descr.txt