Selection of indicators by machine learning: Application to estimate permanent grassland plant richness

Sylvain Plantureux, Jean Villerd, Bernard B. Amiaud, Simon Taugourdeau, Christian C. Bockstaller

Selection of indicators by machine learning: Application to estimate permanent grassland plant richness

Plantureux S.[1], Villerd J.[1], Amiaud B.[1], Taugourdeau S.[1], Bockstaller C.[1]
[1]*UMR Nancy Université - INRA Agronomie et Environnement, Nancy-Colmar, France*

Abstract

Indicator based on key species or other data from field observation are very useful to manage permanent grasslands or to control result-oriented agri-environment schemes. These indicators must generally fulfil the following features: purpose relevance (i.e. optimize forage quality, maximize biodiversity, etc), minimization of the acquisition cost (time and money), stability in time and space, low specialized knowledge requirement, sensitivity. Permanent grasslands are complex ecosystems based on multispecific swards and multiple agronomic and ecological functions. Adequate methodologies are thus required to determine simple indicators.

We tested the regression tree methodology coming from machine learning techniques (artificial intelligence) to predict the plant richness of mountain and lowland permanent grasslands in France. Four potential indicators were considered (alone or combined): species, genus, families, and colours of flowers. Each indicator was associated to its easiness of observation, and the prediction quality of the models was estimated by several criteria (coefficient of determination, relative absolute error). A combination of plant genus and colour of flowers was found as the best compromise in order to estimate permanent grassland plant richness.

Keywords: indicators, richness, biodiversity, machine learning

Introduction

Several countries in Europe have applied output based approaches to agri-environmental management contracts (i.e. MEKA program in Germany, Environmental Stewardship in U.K., Ecological Compensation Area in Switzerland, grassland Agri-environmental measures 2007 in France ...). All these plans require control indicators for administration and farmers themselves. In MEKA program, the result is controlled by the observation of four species, within a list of 28 species elaborated on phytosociological rules (Opperman *et al*., 2003). The ecological and agronomical relevance of this type of criterion has been discussed by Plantureux *et al*.(2010). A key issue for scientists and administration is to find appropriate criteria to control the result, associating scientific relevance and practical feasibility. Machine learning techniques coming from computer science give new opportunities to select rapidly and simply such criteria. The purpose of the present work is to test the regression tree method to provide relevant indicators that predict permanent grassland plant richness.

Materials and methods

The evaluation of the method was conducted in France with a dataset made up of 3792 relevés of permanent grasslands, with a full description of their botanical composition. The grasslands are record in the eFLORAsys database (Plantureux *et al.*, 2010), half of them located in North-eastern part of France and the other ones widespread in the whole country. About 25% of grasslands are in mountainous conditions. Plant richness ranged from 5 to 79 species per grassland (observation area 5-10,000 m^2). We selected the following potential indicators for evaluating plant specific richness (calculated by eFLORAsys): species, genus, flower colours, and combinations of genus and flower colours. In all cases, the presence/absence and not the species dominance in the sward were taken into account. For each indicator we tested the hypothesis that a limited number of values (i.e. a limited number of species or genus) was enough for predicting richness of each of the 3792 grasslands. Unlike the procedure

performed in agri-environmental schemes, the species were not selected by specialists, but found by the computer (machine learning)
We tested the regression tree methodology coming from the machine learning techniques (artificial intelligence) to predict the plant richness. Regression trees aim at predicting the value of a numerical outcome (here, plant richness) with respect to a set of dependant variables (here, presence/absence of the above potential indicators). Their behaviour is similar to decision trees, except that leaves contain linear models of the outcome variable, rather than single values (Quinlan, 1992). We first performed a feature selection pre-processing that extracts a subset of dependant variables by considering the individual predictive ability of each variable along with the redundancy between them (Hall, 1999). Then the regression tree was built and evaluated by cross-validation using Weka software (Hall et al., 2009). The accuracy of the model is measured by the Relative Absolute Error (RAE), the ratio between the sum of errors using the model, and the sum of errors using the mean as a constant predictor. The lower is the RAE, the more accurate is the model. A model with RAE higher than 100% is worse than the mean as a predictor. This method was applied to obtain the smallest list of species to be considered to predict correctly the total plant richness.
A similar procedure was performed for genus and flower colours, but considering combinations of flowering by month (i.e. white, yellow and red in May + white and blue in June + white in September).

Results and discussion
Main results for the potential indicators are presented in table 1. We found 989 plant species within the 3792grasslands, but 264 of them are enough to predict plant richness with a high coefficient of determination (CD=0.98) and a low relative absolute error (RAE=14%).

Table 1: Indicators for the prediction of plant species richness of the 3792 grassland of 4 French Natural Regional Parks. Relative Absolute Error (REA) estimates model predicting quality.

Indicator	Criteria	Value	Coefficient of Determination	Relative Absolute Error
Species (best fit model)	Number of species	264 species	0.98	14%
Species (model with limited number of species)	Number of species	34 species	0.85	37%
Genus (best fit model)	Number of genus	135 genus	0.96	17%
Genus (model with limited number of genus)	Number of genus	43 genus	0.86	30%
Flower colours for each month	Number of combinations flower colour x month	77 combinations	0.86	35%
Flower colours in May	Nb de couleurs	10 colours	0.62	61%
Combination of genus and Flower colours in June	Number of combinations genus x flower colour	41 genus x 5 colours	0.90	29%

The best compromise between the species number (to simplify the previous indicator) and the model quality was found for a model with 34 species (over the 989 species). A list of 135 plant genus is able to predict plant richness, and the CD but not the RAE is almost unaffected by a reduction to 43 genus. 77 combinations of flower colours and months are required to

correctly predict plant richness, but this leads to examine grassland several times from April to September. Flower colours just appearing in a single month cannot be considered as good indicators, the best of them (flower colours in May) presenting a very poor RAE. Finally a combination of genus and flower colours in June is comparable (CD and RAE) to genus, the number of genus to determine just decreasing from 43 to 41.

The recognition of key species appears as a powerful method to predict plant richness, but this indicator requires a specialized knowledge as some retained species belong to the same genus (i.e. *Carex* or *Astragalus)* or are not common (i.e. *Coincya cheiranthos subsp. montana (DC.) Greuter & Burdet* or *Dryopteris carthusiana (Vill.) H.P.Fuchs*). Genus are more easily recognizable by non botanists, and the indicator considering 43 genus (table 2) appears as the best compromise between feasibility and accuracy.

Table 2: List of 43 genus in the best fit model predicting plant richness of grasslands

Achillea- Agrostis- Ajuga- Anthoxanthum- Bellis- Brachypodium- Briza- Bromus- Carex- Carum- Centaurea- Cerastium- Cirsium- Colchicum- Crepis- Cynosurus- Daucus- Equisetum- Euphorbia- Festuca- Filipendula- Galium- Holcus- Koeleria- Lathyrus- Leucanthemum- Lotus- Luzula- Narcissus- Ononis- Ornithopus- Plantago- Primula- Prunella- Rhinanthus- Rumex- Senecio- Silaum- Stellaria- Succisa- Trisetum- Veronica- Vicia

Considering only the colours of flowers is potentially a very simple indicator with a brief acquisition, but results showed either a good prediction requiring repeated observations over the growing season, or a single observation (in May) associated to a poor quality of prediction.

Conclusion

In a methodological point of view, the regression tree methodology is an efficient way to rapidly select qualitative potential indicators to predict quantitative values. The good correlations observed for the best models lead to the conclusion that a rapid and a reliable diagnosis of grassland biodiversity can be done from quite simple indicators (i.e. genus) based on a limited effort of data acquisition.

Aknowledgments: Funding for this project has been supplied by the French Ministry of Ecology (Research program DIVA2 « MAE résultat »). In addition, data were provided by four Natural Regional Parks: Ballons des Vosges, Brenne, Haut-Jura, Massif des Bauges.

References

Hall M.A. (1999) Correlation-based feature selection for machine learning. PhD Thesis, University of Waikato.
Hall M.A., Frank E., Holmes, G., Pfahringer B., Reutemann P., Witten I.H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
Opperman R., Gujer H. (2003). Artenreiches Grünland berwerten und förden – MEKA und ÖKV in der Praxis, Stuttgart, Ulmer.
Plantureux S., Ney A., Amiaud B. (2010). Evaluation of the agronomical and environmental relevance of the CAP measure 'flowered grassland'. Proceedings of the 23th General Meeting of the European Grassland Federation. Kiel (Allemagne) 29 august-2 september 2010, 666-668. ISBN 978-3-86944-021-7
Plantureux S., Amiaud B. (2010). e-FLORA-sys, a website tool to evaluate the agronomical and environmental value of grasslands. Proceedings of the 23th General Meeting of the European Grassland Federation. Kiel (Allemagne) 29 august-2 september 2010, 732-734. ISBN 978-3-86944-021-7
Quinlan R. (1992) Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348