 DOI: 10.3217/978-3-85125-954-4-18

Influence of Data Processing on Hyperspectral-Based Classification of Managed Permanent Grassland

Viktoria Motsch, Roland Britz, Andreas Gronauer
University of Natural Resources and Life Sciences, Vienna,
Department of Sustainable Agricultural Systems, Institute of Agricultural Engineering
Peter-Jordan-Straße 82, 1190 Vienna, Austria
viktoria.motsch@boku.ac.at

Abstract

The botanical composition of grassland stands can be determined using a combination of hyperspectral imaging and machine learning. Data processing before machine learning can significantly improve overall model performance. Specific preprocessing variants, such as smoothening and derivation of the spectrum, were found to be beneficial for classifying grassland species groups in detached models using hyperspectral data from permanent grassland obtained under laboratory conditions. Compared to extensively preprocessed data, raw spectral data yielded no statistically decreased performance in most cases.

1. Introduction

Grassland vegetation typically comprises grasses, herbs, and legumes which represent different functional traits [14] and feed values; knowledge of their relative proportions offers several advantages for site-specific management and livestock feeding. Remote sensing is a non-destructive method used for the reproducible sensing of large areas [16] as detected spectral signatures may vary depending on the species group. Machine learning models based on hyperspectral data can be used for species group classification [4, 5]. For this, data preprocessing might be a substantial step in enhancing model performance. The use of derivatives together with spectral data is a common technique [10,18] as removes background signals and visualizes spectral curve shape differences that might not be evident in the spectra [7]. Smoothing operations such as Savitzky-Golay filtering are frequently applied [6, 8] as well as data standardization or normalization (see Fig. 1). A systematic review under laboratory conditions can reveal the influence of the vast number of data processing variants in combination with machine learning on the spectral-based classification of permanent grassland vegetation.

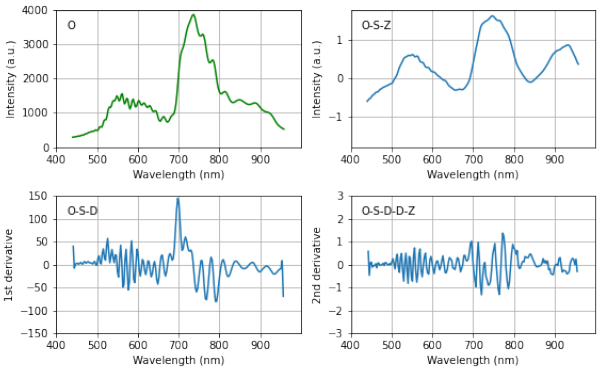

Figure 1. Representative reflectance spectrum and different preprocessing variants for a single red clover (*Trifolium pratense L.*) sample. Left upper corner denotes preprocessing variant.

2. Materials and Methods

The dataset used throughout this study is described in detail by Britz *et al*. [5]. Briefly, an in-house hyperspectral imaging setup was used under standardized laboratory conditions. In total, 5768 plant samples were acquired at two Austrian grassland sites. Each sample was derived from an individual plant, manually annotated and labeled according to species group (grass, herb, or legume).

2.1. Data Preprocessing

For each sample, a total of 100 pixels were drawn randomly stratified. All samples were grouped based on their species group, then randomly stratified and assigned a chunk number from 1 to 5. Further, data was preprocessed using different combinations of Savitzky-Golay-smoothening (function savgol with a filter length of 5 and quadratic filter from R package pracma 2.3.3 [3]), derivation, and Z-standardization (see Tab. 1). In total, 27 preprocessing variants were generated and analyzed.

Step	Variant																								
1.	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
2.		Z	S	S	D	D	D	D	S	S	S	S	D	D	D	D	D	D	D	D	S	S	S	S	S	S
3.			Z		Z	S	S	D	D	D	D	D	D	D	D	S	S	S	S	D	D	D	D	D	D
4.						Z		Z	S	S		Z	S	S	D	D	D	D	D	D	D	D	S	S	S
5.									Z			Z		Z	S	S		Z	S	S	D	D	D		
6.																Z			Z		Z	S			

Table 1. Preprocessing variants generated from original data (O). D = derivative, S = Savitzky–Golay filter, Z = Z-standardization.

2.2. Machine Learning Algorithms

Multi-Layer Perceptron (MLP), Random Forest (RF), and Partial Least Squares Discriminant Analysis (PLS-DA) models were trained for species group classifications. The class weights were normalized to compensate for unbalanced classes. Final training was performed, 5-fold cross-validated, and performance metrics were calculated based on validation parts not used for training. Details on machine learning algorithms can again be found in Britz *et al*. [5].

Briefly, MLP networks were trained using Python, PyTorch [13], Tune [9] included in Ray [12] and hyperopt [2]. The architecture is a fully connected layer followed by batch normalization and a rectified linear unit activation function (ReLU). After another fully connected layer with a ReLU, the final layer is connected to the three output classes. Cross-entropy loss with class weights was used together with a stochastic gradient descent optimizer. Hyperparameters for each variant were searched using an ASHA and in total 100 hyperparameter combinations per dataset variant and group were evaluated. The five hyperparameter combinations, having achieved the highest accuracy per dataset variant and group, were retrained with 5-fold cross-validation for 120 epochs. Then, the model with the highest cross-validated accuracy found at any epoch is depicted in the results. RF classifiers were trained using the function ranger from the ranger package [17] with mtry of 40, SF of 1 and 400 trees, resulting in reasonable accuracy and computation time for training. PLS regression was performed using the cppls function from the pls package [11] with 64 components. Subsequently, linear DAs with the lda function from MASS package [15] were performed.

3. Results and Discussion

MLP achieved cross-validated accuracies of 96.9 % for species group (grass, herb, or legume) classification. While MLP and PLS-DA performed well across a wide range of preprocessing variants and showed a high generalization ability, this was not true for RF (see Fig. 2). The main reason for this is that RF usually uses only a few predictors at the tree level to form a decision boundary [1], which makes it more sensitive to data variations than MLP and PLS-DA.

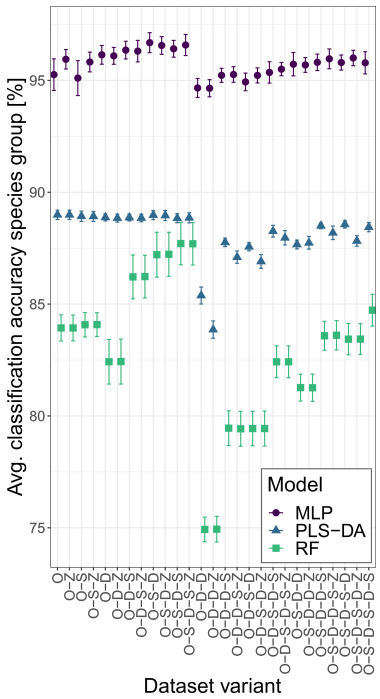

Figure 2. Mean species group classification accuracy based on the preprocessing variant for multilayer perceptron (MLP), partial least squares discriminant analysis (PLS-DA), and random forest (RF) models. X-axis abbreviations (preprocessing steps from bottom to top): O = original data, D = derivative, S = Savitzky–Golay filter, Z = Z-standardization. Error bars indicate standard deviation, 5-fold cross-validated.

In general, similar trends in classification accuracy could be observed depending on the preprocessing variant. Variants differing only in subsequent Z-standardization showed no significant differences independent of model type. Preprocessing steps that do not lead to increased accuracy should be avoided for the sake of simplicity. Preprocessing variants including a Savitzky-Golay filter before a derivation work particularly well for data with low spectral band distances. Here, differences between successive spectral channels may be slight compared to random noise [7]. Other variants can also benefit from Savitzky-Golay filtering as a noise reduction technique. Interesting preprocessing variants that performed well, independent of the model type, included the combination S-D without a second D. In particular for RF but also in other models, variants containing a derivation (D) without prior Savitzky–Golay filter (S) mainly performed worse than variants with a combination of S and D. This underlines the usefulness of spectral gradients in combination with smoothing for machine learning applications. However, for MLP and PLS-DA, even the original dataset variant (O) generated models that were not significantly different from the best statistical model.

References

[1] Houman Abbasian, Chris Drummond, Nathalie Japkowicz, and Stan Matwin. Robustness of Classifiers to Changing Environments. In Atefeh Farzindar and Vlado Kešelj, editors, *Advances in Artificial Intelligence*, pages 232–243, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[2] James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[3] Hans W. Borchers. *pracma: Practical Numerical Math Functions*, 2021. R package version 2.3.3.

[4] Roland Britz, Norbert Barta, Andreas Klingler, Andreas Schaumberger, Alexander Bauer, Erich M Pötsch, Andreas Gronauer, and Viktoria Motsch. Hyperspectral-based classification of managed permanent grassland with multilayer perceptrons: Influence of spectral band count and spectral regions on model performance. *Agriculture*, 12(5):579, 2022.

[5] Roland Britz, Norbert Barta, Andreas Schaumberger, Andreas Klingler, Alexander Bauer, Erich M Pötsch, Andreas Gronauer, and Viktoria Motsch. Spectral-based classification of plant species groups and functional plant parts in managed permanent grassland. *Remote Sensing*, 14(5):1154, 2022.

[6] P. D. Dao, Y. He, and B. Lu. Maximizing the quantitative utility of airborne hyperspectral imagery for studying plant physiology: An optimal sensor exposure setting procedure and empirical line method for atmospheric correction. *International Journal of Applied Earth Observation and Geoinformation*, 77:140–150, May 2019.

[7] Tanvir H. Demetriades-Shah, Michael D. Steven, and Jeremy A. Clark. High resolution derivative spectra in remote sensing. *Remote Sensing of Environment*, 33(1):55–64, jul 1990.

[8] T. Fricke and M. Wachendorf. Combining ultrasonic sward height and spectral signatures to assess the biomass of legume-grass swards. *Computers and Electronics in Agriculture*, 99:236–247, Nov. 2013.

[9] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118*, 2018.

[10] F. Locher, H. Heuwinkel, R. Gutser, and U. Schmidhalter. Development of Near Infrared Reflectance Spectroscopy Calibrations to Estimate Legume Content of Multispecies Legume-Grass Mixtures. *Agronomy Journal*, 97(1):11–17, Jan. 2005.

[11] Bjørn-Helge Mevik and Ron Wehrens. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2):1–23, 2007.

[12] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A Distributed Framework for Emerging AI Applications. *CoRR*, abs/1712.05889, 2017.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] J. Schellberg and L. da S. Pontes. Plant functional traits and nutrient gradients on grassland. In *16th Symposium of the European Grassland Federation "Grassland Farming and Land Management Systems in Mountainous Regions"*, volume 16, pages 470–483, Gumpenstein, Austria, Aug. 2011. Grassland Science in Europe.

[15] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag GmbH, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[16] M. Wachendorf, T. Fricke, and T. Möckel. Remote sensing as a tool to assess botanical composition, structure, quantity and quality of temperate grasslands. *Grass and Forage Science*, 73(1):1–14, Mar. 2018.

[17] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[18] H. Yu, B. Kong, G. Wang, H. Sun, and L. Wang. Hyperspectral database prediction of ecological characteristics for grass species of alpine grasslands. *The Rangeland Journal*, 40(1):19–29, 2018.