FEATURE FILTERING AND SELECTION FOR DRY MATTER ESTIMATION ON PERENNIAL RYEGRASS: A CASE STUDY OF VEGETATION INDICES.

G.T. Alckmin [1,2*], L. Kooistra [2], A. Lucieer [1], R. Rawnsley [1,3].

[1] School of Technology, Environments and Design, University of Tasmania, Hobart, Australia - (gustavo.alckmin, arko.lucieer, richard.rawnsley)@utas.edu.au
[2] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University and Research, P.O. Box 47, 6700 AA Wageningen, The Netherlands - lammert.kooistra@wur.nl
[3] Tasmanian Institute of Agriculture, University of Tasmania, Private Bag 3523, Burnie 7320, Tasmania, Australia

KEY WORDS: Feature Selection, Collinearity, Vegetation Indices, Biomass, Dry Matter, Pasture, Perennial Ryegrass, Machine Learning.

ABSTRACT:

Vegetation indices (VIs) have been extensively employed as a feature for dry matter (DM) estimation. During the past five decades more than a hundred vegetation indices have been proposed. Inevitably, the selection of the optimal index or subset of indices is not trivial nor obvious. This study, performed on a year-round observation of perennial ryegrass (n = 900), indicates that for this response variable (i.e. kg DM.ha^{-1}), more than 80% of indices present a high degree of collinearity (correlation > |0.8|.) Additionally, the absence of an established workflow for feature selection and modelling is a handicap when trying to establish meaningful relations between spectral data and biophysical/biochemical features. Within this case study, an unsupervised and supervised filtering process is proposed to an initial dataset of 97 VIs. This research analyses the effects of the proposed filtering and feature selection process to the overall stability of final models. Consequently, this analysis provides a straightforward framework to filter and select VIs. This approach was able to provide a reduced feature set for a robust model and to quantify trade-offs between optimal models (i.e. lowest root mean square error - RMSE = 412.27 kg.DM.ha^{-1}) and tolerable models (with a smaller number of features - 4 VIs and within 10% of the lowest RMSE.)

1. INTRODUCTION

Vegetation indices have been extensively employed on precision agriculture (PA) and remote sensing (RS) to estimate different biochemical and biophysical attributes of vegetation. VIs can be portrayed as an example of feature engineering in RS, in which the combination of different bands outputs a new feature that should display higher explanatory value. Such technique has been employed since the infancy of RS through simple ratios (Jordan, 1969) ensued by normalized differences (Rouse, Hass, Schell, & Deering, 1973). During the past decades, over a hundred vegetation indices have been proposed (Xue & Su, 2017), surpassing the possibility of optimally or accurately employing each one of them without extensive domain knowledge. This knowledge is not, however, universally available nor readily deployable.

A characteristic of hyperspectral data from vegetation is the high level of multicollinearity within portions of the spectrum. In practical terms, such redundant and numerous features will result in models that require a substantially larger number of observations and time to be confidently trained/validated (*Hughe's phenomenon*). More importantly, this may result in models which are prone to overfit (ergo, not perform well in unseen datasets.) To attenuate such issue, there are, mainly, four different approaches to this problem: (i) filtering, (ii) feature engineering (exemplified earlier), (iii) feature extraction and (iv) feature selection.

Feature extraction is usually performed through transformation methods such as principal component analysis (PCA) or partial least squares (PLS). A major issue with these methods is that principal components and latent variables are, in most cases, not interpretable as these are linear combinations of the original feature space.

Alternatively, filtering and variable selection can be coupled to (a) decrease the level of multicollinearity, (b) take advantage of feature engineering (i.e. vegetation indices/decades of expert knowledge), (c) determine which are the best combination of VIs and (d) reduce the number of bands needed for the accurate estimation.

Furthermore, the identification of a small set of VIs which yield a generalizable model is an important design parameter when developing custom-made multispectral sensors.

In summary, the objective of this study is, therefore, to develop a framework based on filtering and selection of VIs that could provide a feature-space with several positive characteristics: condensed expert knowledge, fewer bands, less obvious multicollinear features and, ultimately, a robust model with a minimum number of features.

* Corresponding author

2. MATERIALS AND METHODS

2.1 Data Collection:

The experiment was established at the Tasmanian Dairy Research Farm in Elliot, TAS (Australia - 41°04'57.3"S 145°46'21.8"E.) The experimental layout can be described as a set of 30 plots (1.3 x 7.5m) of perennial ryegrass (*Lolium perenne*). The plots were grouped in three main blocks (10 plots per block); each block was split in two different growth intervals (long and short or approximately 30 and 15 days, respectively), each plot on the split-block was randomly allocated a different fertilizing regime (0, 25, 50, 75 or 100 kg of N). A schematic layout is presented on Figure 1 .

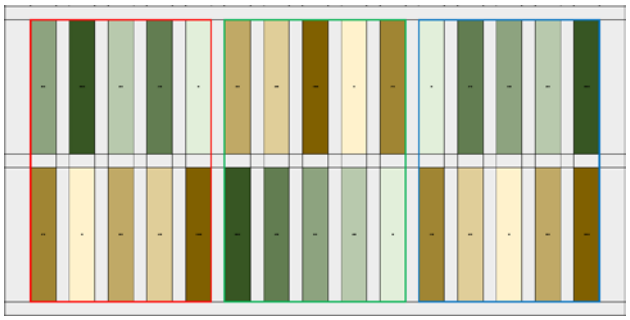

Figure 1- Perennial ryegrass plot layout: colours and shades are linked to growth interval and fertilization level, respectively.

Spectral data was collected by an ASD Handheld 2 (Colorado, USA) on five different dates, always around solar noon and under clear-sky conditions: December 2016, February, April, October and November 2017.

This instrument acquires data from 325 to 1075 nm, with a total of 750 bands and field of view equal to 25º. Total time spent to acquire all measurements (180 samples) was on the range of 1.5 to 2 hours. The sequence of measured plots was randomized to minimize any systematic effect of solar position across the plots during data collection.

The instrument setup was as follows and as per the manufacturer's recommendation: 30 scans for spectrum averaging; 60 scans for dark current and white reference.

Within each plot, six randomly allocated sample-sites were selected. Spectral measurements (i.e. samples) were taken at approximately one-meter height; thus, with a footprint equal to 0.15 m² (or 0.44 m diameter). Each sample-site was measured five times; the final sample spectral data was the average value of these five measurements.

In addition, after finishing measuring the samples of each plot, a scan of the white reference (Spectralon®) was recorded. The intention of this procedure was two-folded: (a) to monitor the stability of the instrument and (b) detect any possible change in atmospheric conditions. Additional best practices were in accordance to the instruction of the user's manual.

The sensor footprint was manually harvested (up to residue-height of approximately five cm), stored on micro perforated plastic baking bags and dried for 48 hours at 60 °C on a forced air oven immediately after harvest. Afterwards, weights were measured on a digital scale (MassCal, 30kg ±0.5g). Consequently, there are 180 observations per campaign and, thus, 900 observations total.

2.2 Data Analysis - Feature Filtering and Selection:

The framework for data analysis (Figure 2) consists of three different steps: (A) filtering of highly correlated and non-significant features; (B) recursive feature elimination and feature selection (best and tolerable subsets) and (C) model validation.

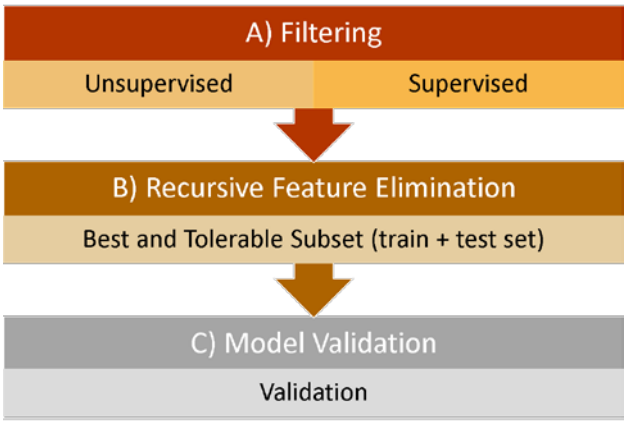

Figure 2 - Framework for data analysis

2.2.1 Filtering: Spectral data (reflectance) was transformed in 97 different VIs (package *hsdar::vegindex*). Subsequently, a correlogram (package *caret::findCorrelation)* was created by which a maximum cut-off can be applied to eliminate highly correlated features (unsupervised filtering). Features above the cut-off are evaluated in a pair-wise fashion: the one with the largest mean correlation (i.e. correlation with all other features) is removed.

As an additional filtering, a minimum cut-off equal to |0.2| (arbitrary value) between the remaining filtered features and kg.DM.ha^{-1} values (supervised filtering) was applied.

2.2.2 Recursive Feature Elimination: after the filtering process, a recursive feature elimination (package *caret::rfe*) is performed against the training set (70% observations, n = 630, random forest regression, repeated cross-validation – 5 repeats, 10 folds). Ranking of features, as well as variable importance, are calculated through a random-forest routine. Within such analysis, the optimal (subset of features with minimum RMSE) and tolerable models are identified. The tolerable feature set is the smallest group of variables which have presented (in training-testing stages) results which are less than 10% (arbitrary value) above of the minimum RMSE (optimal) model. Such process is repeated for each of the five different correlation cut-off levels.

2.2.3 Model Training and Validation: The variable importance for each run (5 repeats, 10 folds) at the tolerable number of features is extracted. The highest-ranking features (variable importance) are then used as input to train a random-forest model on the training/testing dataset (n = 630). The trained model performance is assessed against the validation dataset (n = 270).

2.2.4 Analysis of Filtering and Feature Selection:

As previously stated, ideally, this framework should reduce the size of the feature-space without compromising the level of accuracy/model performance (e.g. RMSE and/or r-squared) in a validation dataset. To assess this hypothesis, five different correlation cut-off levels are tested: 99%, 95%, 90%, 85% and 80%. Such analysis serves three additional purposes: (a) check whether the number of variables in a tolerable feature-space changes, (b) which VIs are selected according to the cut-off thresholds and (c) if these selected features are the same after a specific threshold. In other words, if the feature filtering and selection process stabilizes and robust/generalizable models are created.

3. RESULTS

Overall, the different growth periods and fertilization regimes were able to provide a wide gradient of DM, with values ranging from 111.8 kg.ha^{-1} (minimum) to 4662.8 kg.ha^{-1} (maximum) and a mean equal to 1647.8 kg.ha^{-1}. Given the different seasons and weather conditions, the dataset should provide an adequate surrogate for on-farm pasture conditions.

As initially discussed, VIs are highly correlated and the process of filtering drastically decrease the number of remaining features as shown in Figure 3. However, this has not considerably affected model performance for models with more than three or four VIs, as shown in Figure 4.

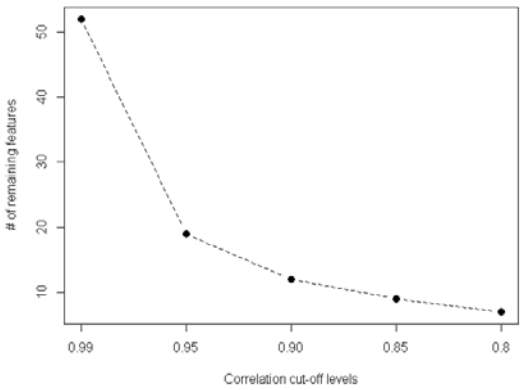

Figure 3 - Number of remaining features by level of correlation after filtering process.

There is, however, a noticeable difference when trying to select a model with only one feature. In this case, the cut-off threshold and random-forest routine may result in noticeable differences, with RMSE differences ranging of around 100 kg.DM.ha^{-1} (Figure 4).

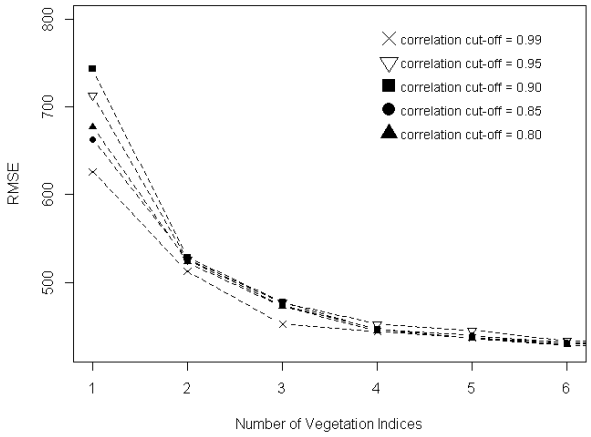

Figure 4 - Root-mean-square (RMSE) error for different feature filtering and selection cut-off levels.

The process of selecting a tolerable model (black-triangle), rather than the optimal model (black-circle) also highly decreases the number of input features (Figure 5). In the case of a correlation cut-off threshold of |0.99|, the number of variables used in the tolerable model is equal to four, whereas the optimal model has 45 features. On interest of avoiding similar Figures, only the selection process of tolerable/optimal model for a cut-off of 0.99 is displayed.

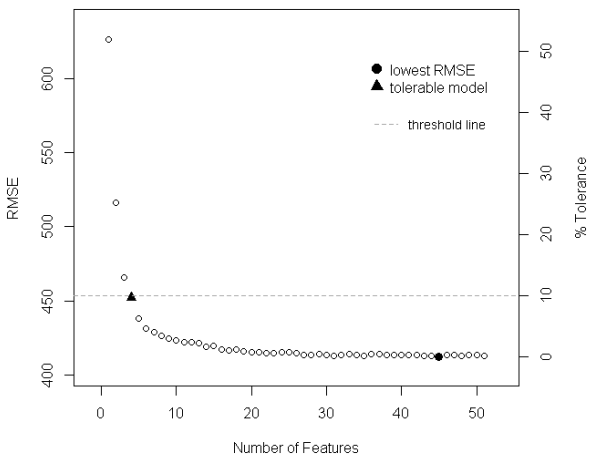

Figure 5 - Optimal (circle) and Tolerable (triangle) Model Selection

For the remaining correlation cut-off thresholds, the number of features on a tolerable model is also equal to four. It is important to stress, however, that (as per Figure 4) model performance on training/test stages for all initial cut-off thresholds (and four variables) is similar (i.e. RMSE ≈ 430 kg DM.ha^{-1}).

The variables selected for each tolerable model, along as its variable importance (extracted from training and test stages) are presented on Figure 6 and Table 1.

	Filtering Cut-off Level				
Ranking	0.99	0.95	0.90	0.85	0.80
1st	MPRI	CI	CI	CI	CI
2nd	RDVI	Datt3	DDn	SPVI	SPVI
3rd	REP_Li	DDn	Datt3	Datt3	Datt3
4th	SR8	Carter	Carter	Carter	Carter

Table 1 - Selected features by different initial correlation cut-off levels.

It is important to point out that when the correlation cut-off is equal to 0.99 (lowest filtering level), the selected variables are all different from the remaining cut-off models (Table 1.) Equally, the values of variable importance are lower than for the remaining models (Figure 6.)

Noticeably, after the cut-off of |0.95|, the selection of variables stabilizes. As shown on Figure 6, the following VI are selected: CI (Zarco-Tejada, Pushnik, Dobrowski, & Ustin, 2003), Datt3 (Datt, 1999), and Carter (Carter, 1994) (cut-off |0.80| is not displayed as it has the same VIs as |0.85|)

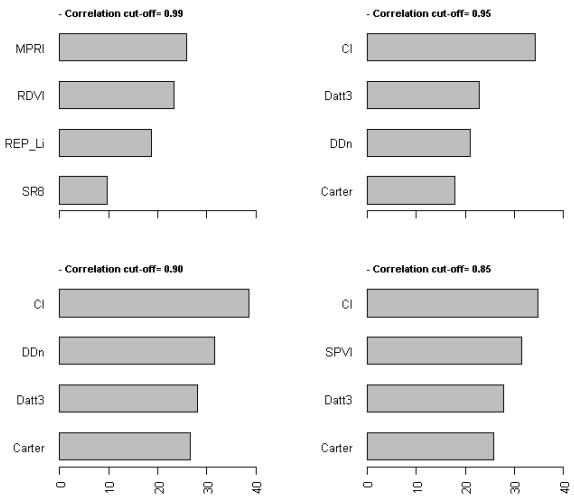

Figure 6 - Variable importance for tolerable dataset

Tolerable models for all five cut-off values are validated (as per Figure 7) and have present similar performance as in training-testing stage. Results are summarized on Table 2: On interest of avoiding similar figures, only the selection process of tolerable/optimal model for a cut-off of |0.99| is displayed.

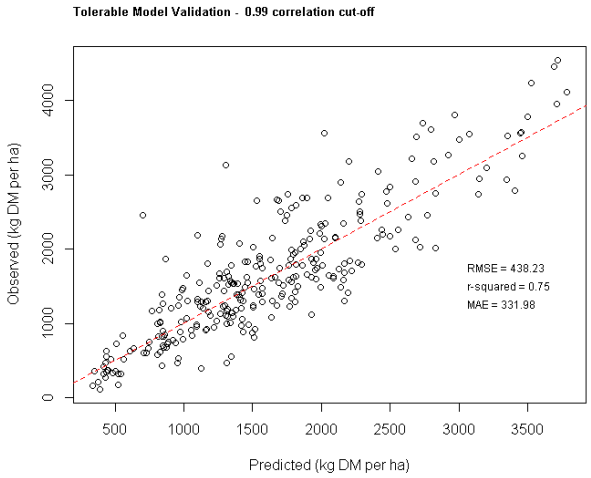

Figure 7 - Model performance (validation). Cut-off = |0.99|

Cut-off	RMSE	r-squared	MAE	Features
0.99	438	0.75	332	4
0.95	436	0.76	328	4
0.90	435	0.76	327	4
0.85	438	0.75	330	4
0.80	438	0.75	330	4

Table 2 - Validation results by cut-off level.

4. CONCLUSION

Overall, this framework was successful to decrease the size of the feature-space, without compromising the level of accuracy of final (i.e. both optimal and tolerable) models.

Largely, whether employing more or less restrictive filtering, tolerable models presented similar performances (Table 2.) In other words, regardless the filtering threshold levels, all models presented similar performance when including three or more features as seen on Figure 4 .

However, when analysing the less-rigid threshold (i.e. |0.99|), variable selection is vastly different from all other models and variable importance is lower than for the remaining models (Figure 6.) Alternatively, these remaining models (i.e. more restrictive filtering models) have consistently selected CI, Datt3 and Carter. Even when applying the most restrictive cut-off threshold (i.e. |0.80|) the pool of seven remaining variables (Figure 8) contain the VIs which are found to be of highest variable importance and would be, subsequently, selected for the tolerable models.

Such VIs are not strongly correlated (Figure 8); thus, avoiding obvious multicollinearity issues and present higher values for variable importance. Such indicates that, in this context, a restrictive filtering can be employed as a simple, yet, powerful initial technique within the task of feature selection.

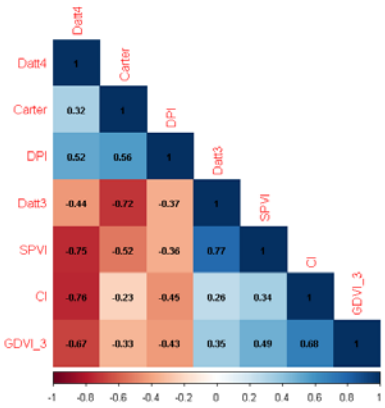

Figure 8 - Correlation of filtered features at cut-off = 0.80

Additionally, this framework was successful to also greatly reduce the number of features while providing the possibility to evaluate the trade-off between subsets of the initially filtered variables. In short, the discretionary choice for tolerable rather than optimal models can also largely decrease the number of necessary variables (Figure 5).

It is important to stress that after four variables, the addition of new variables has marginal returns on terms of predicting power regardless of the initial filtering cut-off, as per Figure 4.

Most importantly, tolerable models (with only four VIs) were able to generalize, providing comparable results to the train/test set against a validation/unseen dataset.

This indicates a satisfactory approach to estimate a wide range of DM values throughout the year with a small number of VIs.

An additional remark should be made: the plateau of 430 kg.DM.ha^{-1} is equivalent to approximately 6.5g per sample area (0.15 m^2); thus, it is important to also stress that a fraction of this error may well be due the sampling/harvesting technique or even a drying process that, despite our best efforts, has some random noise/error. It seems reasonable, therefore, to make use of the tolerable models as they model the bulk of the phenomena, rather than noise due random effects.

ACKNOWLEDGEMENTS

This research was supported by Dairy Australia, through the Dairy on PAR action. All analysis was performed in R 3.5.1 (R Development Core Team, 2008) , using the *hsdar* (Lehnert et al., 2018) and *caret* (Kuhn, 2008) packages, without which this research would not be possible.

REFERENCES

Carter, G. A. (1994). Ratios of leaf reflectances in narrow wavebands as indicators of plant stress. *International Journal of Remote Sensing*, *15*(3), 517–520. https://doi.org/10.1080/01431169408954109

Datt, B. (1999). Visible/near infrared reflectance and chlorophyll content in eucalyptus leaves. *International Journal of Remote Sensing*, *20*(14), 2741–2759. https://doi.org/10.1080/014311699211778

Jordan, C. F. (1969). Derivation of Leaf-Area Index from Quality of Light on the Forest Floor. *Ecology*, *50*(4), 663–666. https://doi.org/10.2307/1936256

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, *28*(5), 1–26. https://doi.org/10.1053/j.sodo.2009.03.002

Lehnert, L. W., Meyer, H., Obermeier, W. A., Silva, B., Regeling, B., & Bendix, J. (2018). Hyperspectral Data Analysis in R: the hsdar Package. Retrieved from http://arxiv.org/abs/1805.05090

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org

Rouse, J. W., Hass, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the great plains with ERTS. *Third Earth Resources Technology Satellite (ERTS) Symposium*, *1*, 309–317. https://doi.org/citeulike-article-id:12009708

Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, *2017*. https://doi.org/10.1155/2017/1353691

Zarco-Tejada, P. J., Pushnik, J. C., Dobrowski, S., & Ustin, S. L. (2003). Steady-state chlorophyll a fluorescence detection from canopy derivative reflectance and double-peak red-edge effects. *Remote Sensing of Environment*, *84*(2), 283–294. https://doi.org/10.1016/S0034-4257(02)00113-X