

# Use of pasture botanical composition data to improve the accuracy of satellite pasture biomass estimates

Grant P.S. ANDERSON<sup>1\*</sup>, Mitchell RAWLINGS<sup>1</sup>, Zoe LUNNISS<sup>1</sup>, Lorna McNAUGHTON<sup>1</sup>, Laura ROSSI<sup>2</sup>, Cáthal WIMS<sup>2</sup>, Chris ROACH<sup>2</sup> and Cameron I. LUDEMANN<sup>2</sup>

<sup>1</sup>Livestock Improvement Corporation, Private Bag 3016, Hamilton, New Zealand

<sup>2</sup>DairyNZ Ltd, Private Bag 3221, Hamilton, New Zealand

\*Corresponding author: grant.anderson@lic.co.nz

## Abstract

Pasture biomass estimates are valuable to farmers, and satellite pasture biomass estimates can potentially provide the required data for less time and labour. However, the accuracy of satellite estimates of pasture biomass can be affected by the botanical composition of the pasture. In this study, botanical composition data were combined in linear regression models and a general additive model with on-farm and satellite pasture biomass estimates to quantify the increase in predictive power from including botanical composition data. The inclusion of botanical composition data improved the accuracy (model  $R^2$ ) of the satellite pasture biomass estimation; the smallest increase was 0.035 (from 0.725 to 0.760) and the largest increase was 0.111 (from 0.599 to 0.710).

Improving the accuracy of satellite estimations of pasture biomass will allow farmers to make more timely and accurate grazing management decisions.

**Keywords:** NDVI, pasture management, stepwise linear regression, generalised additive model

## Introduction

Monitoring of pasture biomass is necessary for farmers to make timely tactical grazing management decisions (Neal et al. 2017). Traditional methods for measuring pasture biomass, such as rising plate meters and the C-DAX Pasture Meter, are time-consuming, labour intensive and subject to operator error (Chapman et al. 2014). Satellite-based monitoring methods that utilise spectral vegetation indices can provide extensive spatial coverage data for farmers (Ali et al. 2016). One widely used index is the Normalised Difference Vegetation Index (NDVI; NASA 2000). However, these indices often vary in the accuracy of their pasture biomass estimates (Anderson & McNaughton 2018). They also suffer from a saturation effect at high biomass levels (Gu et al. 2013).

The botanical composition of pasture is one possible source of variation in accuracy of biomass measurements. Perennial ryegrass (*Lolium perenne*) and white clover (*Trifolium repens*) are the most common species in New Zealand dairy pastures, but not

the only species, and the proportions of ryegrass and white clover may vary temporally and spatially within a paddock (Chapman et al. 1996). Steltzer and Welker (2006) demonstrated that the ability to predict leaf area index (LAI) from NDVI is affected by botanical composition, which suggests that the ability to predict pasture biomass could be similarly affected.

If botanical composition has a consistent effect on satellite pasture estimates, then including botanical composition information in the estimation model should show a concomitant improvement in model accuracy. If such an improvement were sufficiently great, it would motivate an investigation into the frequency and precision of botanical composition data required to produce this improvement.

The aim of this investigation was to determine the effect of the inclusion of botanical composition on the accuracy of satellite estimates of pasture biomass, using NDVI and a range of models based on mean reflectance values.

## Materials and Methods

### Study location and overall botanical composition

In total, sixty 0.5-ha paddocks on the DairyNZ Ltd Scott Farm (Hamilton; 37.460S, 175.220E; elevation 40 metres above sea level) were used in this study for the collection of botanical composition and pasture biomass data. Data were collected on 107 dates between October 2017 and January 2019. The pastures used were predominantly perennial ryegrass–white clover, on Matangi silt loam and Horotiu silt loam soil types. The pastures were sown in autumn 2016, 2017 and 2018.

### Terrestrial data collection

The biomass (kg DM/ha) of pasture in each paddock was estimated by the same two people by visual assessment undertaken during a weekly farm walk similar to the method described by O'Donovan et al. (2002). In all, 4200 paddock estimates of pasture biomass were provided.

Botanical composition was estimated seasonally (August, October, November, January and April) for each paddock, as follows. Directly before grazing,

representative pasture herbage samples were collected. The samples were collected from four 10 m by 1.5 m harvester cuts within each paddock, above a cutting height of 4 cm. Herbage samples were transferred to a laboratory then dissected into perennial ryegrass leaf, perennial ryegrass reproductive stem, annual ryegrass (*Lolium multiflorum*), white clover, broad-leaved weeds, other grass species and dead material of all species and then oven-dried separately at 95°C to a constant weight (this process took approximately 48 hours). Botanical composition was highly variable, but perennial ryegrass leaf was the predominant component (Table 1). A total of 355 botanical composition records were collected.

Satellite imagery collection

All available satellite images were obtained from Planet Labs’ constellation of Dove satellites (Planet Labs 2015) for the same time period (456 days). We received images for 199 days; we were not sent any images that were heavily overcast or impacted by satellite malfunctions. The images were corrected by Planet Labs for surface reflectance (Planet Labs 2018) to remove atmospheric artefacts, making the images more comparable across time, and across different satellites. This resulted in 304 usable images, as there were 39 days with two satellite images, and 9 with three images. The resolution of the Dove satellites is 3.125 m, or approximately 878 pixels per hectare, and the spectral bands of their imagery comprise blue (455–515 nm), green (500–590 nm), red (590–670 nm) and near-infrared (NIR, 780–860 nm)

The full dataset contained 14,438 records. Each record contained mean values of red, green and blue (RGB) and NIR reflectance for a single paddock for a single day, as well as a timestamp, an automated estimate of cloud cover, and the identifier of the satellite from which the image was obtained.

Table 1 Range of pasture component data.

Pasture component	Proportion of pasture component		
	Mean	Minimum	Maximum
Perennial ryegrass			
– Leaf	0.68	0.16	0.98
Perennial ryegrass			
– Reproductive Stem	0.07	0.00	0.39
Annual ryegrass	0.02	0.00	0.38
White clover	0.13	0.00	0.79
Broad-leaved weeds	0.01	0.00	0.30
Dead material	0.05	0.00	0.33
Other grass species	0.05	0.00	0.39

Dataset construction

Each satellite measurement for a paddock that occurred within two days of a visual biomass measurement for that paddock was associated with the corresponding biomass measurement. When this process was complete, the dataset consisted of 2868 matched records. The records in this dataset were then combined with botanical composition records. For each paddock, biomass/satellite records that occurred within 20 days of a botanical composition record were matched with that record. Range limits of 10 and 15 days were also investigated (data not shown), but results in the subsequent analysis were similar, so the selection that provided the largest dataset was used. This resulted in a dataset with 407 records.

Sixty-seven records with biomass measurements greater than 3000 kg DM/ha were removed from the dataset, so that the well-known NDVI saturation effect did not confound the investigation. As a sense check on the data, NDVI was calculated for each image, and compared with the measured biomass. When extreme outliers in this relationship were observed, the corresponding images were visually inspected, and in all cases were found to be due to cloud in the image that had not been detected by the automated system, or to the satellite estimate and biomass measurement taking place each side of a grazing event (so that the one took place pre-grazing and the other post-grazing). These data points were also removed. This process resulted in a dataset of 245 records with visual biomass measurements, satellite data, and botanical composition data. Just over half (53%) of these data points occurred between October 2018 and January 2019; the remainder were approximately evenly distributed across the period October 2017 to September 2018.

Statistical analyses

All analyses were conducted using R version 3.5.1 (R Core Team 2018). NDVI values were computed for the satellite records of each paddock as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

where NIR is near-infrared reflectance and Red is reflectance in the red band.

The following variables were used in the creation of linear models with respect to the measured biomass: NDVI, red reflectance, blue reflectance, green reflectance, NIR reflectance, perennial ryegrass leaf proportion (RL), perennial ryegrass reproductive stem proportion (RS), annual ryegrass proportion (AR), white clover proportion (WC), broad-leaved weeds proportion (W), other grass species (O) and dead vegetative matter of all species (D); the botanical composition data was taken from the nearest sample up

to 20 days before or after the satellite measurement.

A Generalised Additive Model (GAM) from the R package *mgcv* (Wood 2011) was also used to investigate the relationship between measured biomass and the spectral and botanical composition data. GAMs are a class of statistical model used to reveal and estimate non-linear effects. GAMs replace the linear relationship between predictors and response with non-linear smooth functions. The smooth functions can capture both linear and non-linear relationships.

The models were evaluated using adjusted  $R^2$ , and two information criterion metrics: the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). As the  $R^2$  increases with the number of parameters in the model, it is not useful for comparing models where the number of variables differ. Hence, the adjusted  $R^2$  is used:

$$R^2_{adj} = 1 - [(1 - R^2)(n - 1)/(n - k - 1)]$$

Where  $n$  is the number of points and  $k$  is the number of variables.

The AIC and BIC are calculated as follows:

$$AIC = [-2\log L + 2p]$$

$$BIC = [-2\log L + \log(n)p]$$

Where  $L$  is the likelihood function, and  $p$  is the number of parameters in the model.

### Effect of botanical composition information

Analyses were conducted using stepwise regression (for the linear models) and backward elimination (for the GAM) for variable selection, to determine if the addition of some or all botanical composition variables to spectral data resulted in an improvement

in the accuracy of biomass estimation. Four pairs of models were constructed, with and without botanical composition: a linear model using NDVI alone (NDVI-simple), a linear model using mean reflectance values (red, green, blue, and NIR, hereafter RGBN) with an assumption that the effects of the different variables are additive (RGBN-add), a linear model using RGBN considering interactions between the spectral terms (RGBN-mult), and a GAM using RGBN (RGBN-gam; note that the GAM does not include interactions between the variables).

For each pair of models, ANOVA was used to determine if the gain in explanatory power from including botanical composition data was worth the increase in complexity.

### Results

NDVI had the lowest  $R^2$  value of all four models tested whether or not botanical composition was included in the analysis (Table 2). For all four model pairs (i.e. each model with or without botanical composition), the addition of botanical composition data resulted in an improvement in explanatory power, as shown by the improved adjusted  $R^2$ , and ANOVA significance tests ( $P$ -value  $< 0.001$ ; Table 2). The addition of botanical composition data also resulted in a lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in all cases. The Akaike and Bayesian Information Criteria are measures of model quality, with smaller values indicating models of higher quality. Both take into account the goodness of fit and complexity (number of terms) of a model; they differ in that the BIC penalises complexity more heavily. A model that is too complex for the data it represents can result in overfitting, where it fits too closely to the particular dataset used to build it, and does not generalise well to new data.

**Table 2** Characteristics of biomass prediction models.

Model	Bot. comp*	Predictor variables used	$R^2$ (adj.)	ANOVA significant	AIC	BIC
NDVI-simple	No	NDVI	0.599		3488	3498
	Yes	NDVI + O + WC + RL + AR + W	0.710	Yes	3414	3442
RGBN-add	No	R + G + B + N	0.673		3441	3462
	Yes	R + G + B + N + O + WC + RL + AR + W	0.722	Yes	3405	3440
RGBN-mult	No	R * G * B * N	0.725		3409	3468
	Yes	R * G * B * N + O + WC + RL + AR + W	0.760	Yes	3381	3458
RGBN-gam	No	R + G + B + N	0.721		3413	3476
	Yes	R + G + B + N + RS + D + RL + AR + W	0.792	Yes	3355	3471

\*Bot. comp = Botanical composition data

AR = annual ryegrass proportion; D = dead vegetative matter proportion; O = other grass species proportion; RS = perennial ryegrass reproductive stem proportion RL = perennial ryegrass leaf proportion; W = broad-leaved weeds proportion; WC = white clover proportion

Stepwise selection of terms for inclusion in the model identified annual ryegrass, other grass species, perennial ryegrass leaf, weeds and white clover as explanatory for all the linear models; for the GAM, annual ryegrass and perennial ryegrass leaf were still useful, but dead matter and ryegrass reproductive stem replaced other grass species and white clover as useful variables.

## Discussion

The results obtained here suggest that the addition of botanical composition data can provide a substantial improvement in the accuracy of pasture biomass estimation based on satellite collected images, regardless of the model used. Using only the simple NDVI metric, the  $R^2$  for pasture biomass estimation was 0.599, while the inclusion of botanical composition data increased the  $R^2$  to 0.710. When spectral data (red, green, blue and near-infrared) were used separately in the RGBN-add, RGBN-mult, and RGBN-gam models, the botanical composition data were still useful (with  $R^2$  increases of 0.049, 0.035 and 0.071 respectively). This improvement in each case indicates the value of botanical composition information; even the more sophisticated models still benefited from its inclusion. A clear benefit was gained from treating the satellite wavelengths separately, compared with the simple-NDVI model.

A difference of 10 or greater between two models is generally regarded as strongly significant for both AIC (Burnham & Anderson 2004) and BIC (Raftery 1995). For each model, including botanical composition data decreased the AIC by at least 10. Including botanical composition data decreased the BIC by at least 10 for three of the four models; for RGBN-gam, including pasture botanical composition data decreased the BIC by 5.

The RGBN-gam model with botanical composition data had the highest adjusted  $R^2$  and the lowest AIC, indicating it should be the preferred model based on these criteria. It did not have the lowest BIC; however, the BIC has a greater penalty for additional parameters, and the non-linear RGBN-gam model introduced a large number of parameters. The RGBN-gam model and RGBN-mult model were roughly equivalent in the absence of botanical composition information, but with the addition of the botanical information, the RGBN-gam model showed greater accuracy. This suggests that there are benefits to a non-linear treatment of botanical composition proportion information.

The botanical composition variables used here were proportional values; their use as predictors in the linear models treated them as having a linear effect on biomass estimation regardless of the total biomass present. The observed improvement in the accuracy of satellite

pasture biomass estimation that results from adding untransformed botanical composition variables to the linear models indicates that the proportion variables are a sufficient approximation to allow improved pasture biomass estimates from satellite measurements.

The GAM fits submodels to the botanical composition variables, and results in the best overall model by the metrics of AIC and  $R^2$ , but not by BIC, which is more conservative with respect to increased complexity. The botanical composition variables selected for the GAM, relative to the linear models, include D and RS and exclude O and WC. This selective inclusion could be because the variables already present in the GAM incorporate the explanatory power that would be added by O and WC. The improvement in goodness-of-fit obtained from the GAM ( $R^2$  of 0.792) relative to NDVI-simple and RGBN-add ( $R^2$  of 0.710 and 0.722 respectively) could also indicate the presence of non-linear effects.

The saturation effect that occurs for higher biomass values is a problem for satellite measurements of pasture biomass, particularly during periods where pasture biomass is allowed to accumulate to a high level (e.g. pastures saved for silage production and autumn saved pasture). The dataset for this investigation was restricted to biomass below 3000 kg DM/ha and there is no evidence at this point that the results obtained in this work hold for higher biomass.

All available visual and satellite estimates of pasture biomass were used, with the exception of the high biomass values (greater than 3000 kg DM/ha) which were removed to avoid confounding effects (due to potential satellite image saturation). However, the botanical composition measurements were restricted to the upper range of pasture biomass, as pasture botanical composition was estimated directly before grazing. As botanical composition can vary substantially immediately pre- and post-grazing, this was a possible source of noise in the models. Nonetheless, including the botanical composition data resulted in an improvement in the accuracy of the satellite pasture biomass measurements. Botanical composition was estimated five times per year for each paddock enrolled in the study. Further work is required to determine the trade-offs between the frequency of botanical composition estimation and improvements in the accuracy of the models used to predict pasture biomass from satellite collected images.

The standard error of the residuals in these models ranged from 296 kg DM/ha (NDVI-simple, no botanical composition data) to 213 kg DM/ha (RGBN-gam, with botanical composition data), in proportion to the  $R^2$ .

These results are provided with the caveat that they are based on data from a single farm. Similar analysis on data from a range of locations and botanical

compositions would be required to determine whether these results could be generalised, or a more general model that utilises botanical composition to improve biomass estimation could be found.

### Conclusions/Practical implications/Relevance

On the farm used in this investigation, the inclusion of botanical composition information improved the accuracy of pasture biomass estimation from satellite collected images. If a similar result could be repeated across a range of sites and botanical compositions, then it would be possible to improve the estimation of pasture biomass from satellite imagery by augmenting the satellite data with botanical composition data. Improving the accuracy of pasture biomass estimates would enable users to make timelier and more accurate grazing management decisions.

However, the botanical composition assessment method used in this trial is impractical for use on a commercial dairy farm due to labour requirements and the need for specialised equipment. While this study demonstrates that botanical composition data improves the accuracy of satellite estimates of pasture biomass, we have not determined the frequency or the precision of the botanical composition assessments required to produce this improvement.

### ACKNOWLEDGEMENTS

Thanks to the technical team at DairyNZ Ltd, for collecting the data used in this investigation.

### REFERENCES

- Ali I, Cawkwell F, Dwyer E, Barrett B, Green S. 2016. Satellite remote sensing of grasslands: from observation to management. *Journal of Plant Ecology* 9(6): 649-671.
- Anderson G, McNaughton L. 2018. Validation of a satellite pasture measurement system. In: *Proceedings of the 8th Australasian Dairy Science Symposium*. Palmerston North, New Zealand.
- Burnham KP, Anderson DR. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33(2): 261-304.
- Chapman DF, Parsons AJ, Schwinning S. 1996. Management of clover in grazed pastures: expectations, limitations and opportunities. In: Woodfield DR Ed. *White clover: New Zealand's competitive edge*. Agronomy Society of New Zealand Special Publication No.11/Grassland Research and Practice Series No.6. New Zealand Grassland Association, Lincoln New Zealand. Pp. 55-64.
- Chapman D, McCarthy S, Kay J. 2014. Hidden dollars in grazing management: getting the most profit from your pastures. In: *Proceedings of the South Island Dairy Event*, Invercargill, New Zealand.
- Gu Y, Wylie BK, Howard DM, Phuyal KP, Ji L. 2013. NDVI saturation adjustment: A new approach for improving cropland performance estimates in the Greater Platte River Basin, USA. *Ecological Indicators* 30: 1-6.
- NASA. 2000. Measuring Vegetation (NDVI & EVI). Accessed 8 May 2019 from [https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring\\_vegetation\\_2.php](https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php)
- Neal M, Dela Rue B, Eastwood C. 2017. Defining the value proposition for using technology to improve pasture management and harvest more pasture. In: *Proceedings of the 7th International Tri-Conference for Precision Agriculture 2017*.
- O'Donovan M, Dillon P, Rath M, Stakelum G. 2002. A comparison of four methods of herbage mass estimation. *Irish Journal of Agricultural and Food Research* 41: 17-27.
- Planet Labs. 2018. Planet Surface Reflectance Product. Accessed: 8 May 2019 from [https://assets.planet.com/marketing/PDF/Planet\\_Surface\\_Reflectance\\_Technical\\_White\\_Paper.pdf](https://assets.planet.com/marketing/PDF/Planet_Surface_Reflectance_Technical_White_Paper.pdf)
- Planet Labs Specifications: Spacecraft Operations & Ground Systems 2015. Accessed: 8 May 2019 from <http://content.satimagingcorp.com.s3.amazonaws.com/media/pdf/Dove-PDF-Download>
- Raftery, AE. 1995. Bayesian Model Selection in Social Research. *Sociological Methodology* 25: 111-163.
- R Core Team. 2018. *R: A Language and Environment for Statistical coding*. R Foundation for Statistical Computing, Vienna, Austria.
- Steltzer H., Welker J. 2006. Modeling the effect of photosynthetic vegetation properties on the NDVI-LAI relationship. *Ecology* 87: 2765-2772.
- Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1): 3-36.

