

Biometric characteristics and canopy reflectance association for early-stage sugarcane biomass prediction

Murillo Grespan da Rocha¹, Flávio Margarito Martins de Barros¹, Stanley Robson de Medeiros Oliveira², Lucas Rios do Amaral^{1*}

¹Universidade Estadual de Campinas/FEAGRI, Av. Cândido Rondon, 501 – 13083-875 – Campinas, SP – Brasil.

²Embrapa Informática Agropecuária, Av. André Tosello, 209 – 13083-886 – Campinas, SP – Brasil.

*Corresponding author <lucas.amaral@feagri.unicamp.br>

Edited by: Thomas Kumke

Received August 30, 2017

Accepted March 02, 2018

ABSTRACT: Knowing the spatial variability of sugarcane biomass in the early stages of development may help growers in their management decision-making. Proximal canopy sensing is a promising technology that can identify this variability but is limited to quantifying plant-specific parameters. In this study, we evaluated whether biometric variables integrated with canopy reflectance data can assist in the generation of models for early-stage sugarcane biomass prediction. To substantiate this assertion, four sugarcane-producing fields were measured with an active crop canopy sensor and 30 sampling plots were selected for manually quantifying chlorophyll content, plant height, stalk number and aboveground biomass. We determined that Random Forest and Multiple Linear Regression models are similarly able to predict biomass, and that associating biometric variables such as number of stalks and plant height with reflectance data can assist model performance, depending on the attributes selected. This indicates that, when estimating biomass in the early stages, sugarcane growers can carry out site-specific management in order to increase yield and reduce the use of inputs.

Keywords: Random forest, canopy sensor, vegetation indices, precision farming, data mining

Introduction

Sugarcane is a raw material used in the production of sugar, ethanol and recently, for the production of energy from high-fiber varieties (energy cane). Since its development in the production fields is not uniform (Amaral et al., 2015b), estimating its spatial variability in the early stages can lead to management opportunities to increase yield and reduce the use of inputs.

Sugarcane aboveground biomass is an important attribute to measure, since it is directly linked to the final stalk yield (Sandhu et al., 2012). Given this relationship, the use of canopy sensors presents the possibility of predicting sugarcane stalk yield as a guide for variable-rate nitrogen fertilization (Amaral et al., 2015b; Lofton et al., 2012; Raun et al., 2005). However, although Amaral et al. (2015a) obtained reasonable correlations between canopy sensor measurements and sugarcane biomass in the early stages of experiments conducted under field conditions ($r^2 \sim 0.65$), canopy sensor measurements did not result in reliable biomass estimation throughout the field cycle for crop management purposes. Thus, the integration of other crop information might improve these predictions.

For example, in experiments with rice, Tilly et al. (2015) suggest the applicability of plant height spatial data as a non-destructive estimator of biomass. In several plot-studies on plant-cane, Rossi Neto et al. (2017) identified that the biometric attribute with the greatest influence on crop yield was the plant population. Thus, such data could assist in biomass estimation and guide site-specific management. As a consequence, multiple independent variables must be included in predictive models and multivariate data analysis techniques are required (Abdel-Rahman and Ahmed, 2008; De Benedetto et al., 2013).

In this study we opted to use a machine learning algorithm, Random Forest, and compared its performance to Multiple Linear Regression. The first is robust for a number of problems with training data (e.g. collinearity). It is widely found in the literature having been successfully deployed in several domains, and can capture non-linear relationships (Hastie et al., 2009). The second technique, however, is a linear model and is also one of the most used statistical tools in the literature (Draper and Smith, 1998).

Biometric variables may be a useful supplement to canopy reflectance. Thus, the aim of this paper was to evaluate whether biometric variables in combination with canopy reflectance data can assist in the generation of models for early-stage sugarcane biomass prediction drawing on a comparison of the performance of Random Forest and Multiple Linear Regression.

Materials and Methods

Data sampling

The experiment was conducted in four commercial fields cultivated with sugarcane in the northeastern sector of the state of São Paulo, Brazil (Table 1). A canopy reflectance sensor was used to measure three distinct bands of the spectrum, allowing for the calculation of vegetation indices. According to the crop development variability inferred by the sensor (expressed by the NDVI), 30 plots per field were allocated to five NDVI classes (six randomly selected plots per class), similar to the procedure adopted by Portz et al. (2011). The plots consisted of four sugarcane rows 5 m long spaced 1.5 m apart. For each plot, a buffer of five meters' radius was demarcated and the average reflectance value was obtained for each sensor band as a

Table 1 – Study site characteristics.

Field	Area ha	Year	Variety	Ratoon	Geographic coordinates	Altitude m
1	6.43	2012	SP801816	8°	21°21'46.06" S 48°00'51.38" W	580
2	5.93	2012	CTC 2	4°	21°21'57.12" S 47°58'04.70" W	581
3	6.67	2013	CTC 2	3°	21°22'51.39" S 47°58'33.87" W	590
4	6.39	2013	CTC 2	3°	21°22'49.84" S 47°58'21.80" W	590

function of all the readings from that buffer (Figure 1A and B). The purpose of the procedure was to facilitate correlation of the sensor data with the biometry data sampled in the plots.

The canopy reflectance data were obtained when the crop had reached approximately a stalk height of 0.5 m (Amaral and Molin, 2014) through the Crop Circle ACS-430 active sensor (Holland Scientific Inc., Lincoln, Nebraska, EUA), which emits modulated light and captures the resulting reflected light. In the first year, two independent sensor units (two GNSS receivers, both working with L1 band and internal algorithm correction) were used, whereas in the following year a set with four sensor units was used acting integrally (a single GNSS receiver, similar to the previously used). The readings were collected at 1 Hz from all the crop rows, by the sensor mounted on a high-clearance vehicle and conducted at a machine-travel speed of ~4.44 m s⁻¹, maintaining a constant distance between sensor and plant canopy of approximately 0.8 m. This sensor works with three distinct spectral bands, in the red region (RED, 670 nm), the transition region between red and near infrared (RED-EDGE, 730 nm) and the near infrared region (NIR, 780 nm). Thus, in order to enhance the possibility of assisting in the prediction of sugarcane biomass, vegetation indices were calculated from these reflectance data (Table 2).

On the day after the sensor measurements were taken, the following data were sampled along a linear 20 m of each plot (four rows 5 m long): chlorophyll content, obtained by a chlorophyll meter (SPAD-502, Konica Minolta Sensing Inc., Sakai, Osaka, Japan) on two separated diagnostic leaves (+1 and +3 leaves, first and third leaf fully open, respectively – SPAD+1 and SPAD+3), in the median portion of the leaves, and the plot value corresponded to an average of twenty leaves; plant height, corresponding to the average of three plants, measured from the soil to the base of the leaf +1 (HEIGHT); and stalk number, obtained by counting all tillers that belonged to the plot (STALKN). The aerial part of the plants of three 1.5 m rows of each plot was manually cut and weighed for biomass determination.

Variable selection

The dataset used is composed of 120 instances or observations (thirty plots in four fields) and fourteen attributes, namely: height and number of stalks, chlorophyll content in leaves +1 and +3, individual bands (RED, RED-EDGE and NIR), vegetation indices (NDVI, NDRE,

Table 2 – Vegetation indices used in different approaches.

Index	Equation
Normalized Difference Vegetation Index	NDVI = (NIR-RED) / (NIR + RED)
Normalized Difference Red-Edge Index	NDRE = (NIR-REDGE) / (NIR + REDGE)
Chlorophyll Index	CI = (NIR / REDGE)-1
MERIS Terrestrial Chlorophyll Index	MTCI = (NIR-REDGE) / (REDGE-RED)
Soil-adjusted Vegetation Index	SAVI = (1 + 0.5)*(NIR-RED) / (NIR + RED + 0.5)
Canopy Content Chlorophyll Index	CCCI = NDRE / NDVI

RED = reflectance of red region (670 nm); REDGE = reflectance in the transition region between red and near infrared (730 nm); NIR = reflectance in the near infrared region (780 nm).

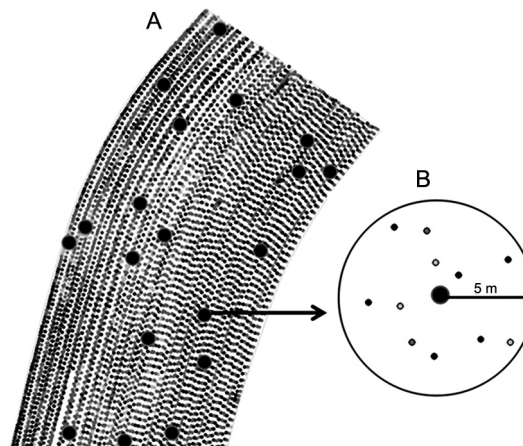


Figure 1 – Part of a field showing sensor data (small dots) and allocated plots (large black dots) according to five vegetation index classes (A) and, highlighted, a 5 m buffer selecting points to represent a plot (B).

CI, MTCI, CCCI and SAVI), and biomass as the meta-attribute (response variable). Since the proposal was to create multivariable models, it was fundamental to an understanding of which attributes should be considered. Several methods can be used when making this selection, and often the subjectivity of the scientist must also be taken into account. Thus, three different approaches to

attribute selection (Approaches 1, 2 and 3 - Table 3) were evaluated and compared to univariate linear regressions. One of the main goals for this selection was the removal of highly correlated variables (collinearity). When adjusting a prediction model using a set of highly correlated variables, the quality of the model can decline, since they do not add information to the model. On the contrary, they can add noise (Draper and Smith, 1998). Consequently, attribute selection facilitates data visualization, reduces storage requirements as well as processing time and improves model efficiency (Guyon and Elisseeff, 2003). Therefore, all the correlations were analyzed, two by two (24 correlations), excluding those variables which presented correlation greater than 0.7 with the others (Figure 2). However, where the relationship was not linear, we chose to keep both variables to allow for possible non-linear behavior in the prediction models.

In Approach 1, the NDRE index was used rather than NDVI because Amaral et al. (2015a) showed that it

has a superior relationship with the sugarcane biomass. Thus, by evaluating subsequent correlations, we chose to retain the attributes HEIGHT, REDGE and CCCI ($r < 0.7$); the attribute NIR (even with $r = 0.87$) was not excluded due to non-linear trends presented.

Using the same principle, to evaluate the possibility of using the NDVI (one of the most used indexes in the literature) instead of NDRE, it was necessary to change HEIGHT by STALKN since NDVI and HEIGHT were highly correlated (Approach 2).

Although correlation analysis is one of the main forms of attribute selection, an approach in which the morph-physiological data were complementary was also adopted (Approach 3). For this approach, we decided not to use separated bands, since the information was already included in the vegetation indices and, thus, focused on the other biometric variables. We therefore decided to keep STALKN and HEIGHT in all situations. Furthermore, we investigated chlorophyll readings in different leaves; in this case, SPAD in leaf +3 was better than SPAD +1 in all situations, while removing SPAD readings from the analysis impaired the simulated models (data not shown). All vegetation indices were tested in order to create new analysis possibilities, but the models that used NDVI prevailed over the others (data not shown; approach 3 - Table 3).

Table 3 – Attributes maintained according to the three attribute selection approaches, while biomass is the response variable.

Approach	Attributes				
1	HEIGHT	REDGE	NIR	NDRE	CCCI
2	STALKN	REDGE	NIR	NDVI	CCCI
3	STALKN	HEIGHT	SPAD+3	NDVI	-

HEIGHT = plant height in the sampling spot; STALKN = stalk/stem quantity in the sampling spot; SPAD+3 = chlorophyll content, obtained by a chlorophyll meter in the leaf +3; REDGE = reflectance in the transition region between red and near infrared (730 nm); NIR = reflectance in the near infrared region (780 nm); NDVI = Normalized Difference Vegetation Index; NDRE = Normalized Difference Red-Edge Index; CCCI = Canopy Content Chlorophyll Index.

Modeling

In this study we opted for using a machine learning algorithm, Random Forest, and compared its performance to Multiple Linear Regression. The first is robust to a number of problems with training data (e.g. col-

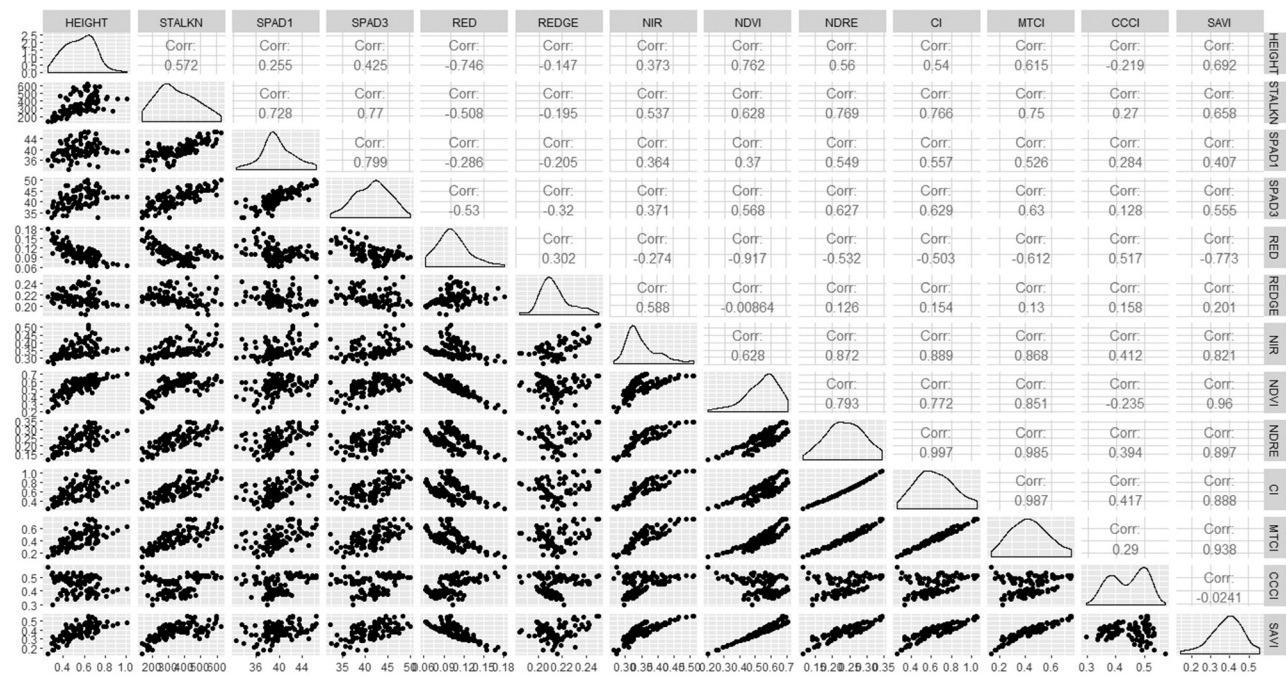


Figure 2 – Correlation between variables and their linearity; at principal diagonal, data distribution (histogram) of each variable.

linearity), is widely found in the literature, has been successful in several domains, and captures non-linear relationships (Hastie et al., 2009). It is an ensemble learning method for classification or regression, that operates by constructing a multitude of decision trees at training time and outputs the most popular class which is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Han et al., 2011). The second technique, however, is a linear model and one of the most used statistical tools in the literature (Draper and Smith, 1998). It involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, whereby more than two attributes are involved and the data are fit to a multidimensional surface.

To do so, scripts were developed with the R Project for Statistical Computing. In particular, we used the Caret package (Classification and Regression Training), which is a compilation of functions that aims to facilitate the creation of optimized models, since it uses internal mechanisms to adjust hyperparameters (Kuhn, 2010).

Considering that the dataset consisted of 120 observations, cross-validation was used (k-fold = 10) for model adjustments, since this approach is recommended for small datasets, including simple (univariate) regressions performed to compare the multivariable models' performance. The metrics used for such comparison were the multiple determination coefficient (R^2), root mean squared error (RMSE), and Ratio Percentage Deviation (RPD - ratio between the actual biomass standard deviation and the RMSE of the validation). According to the classification proposed by Viscarra Rossel et al. (2006), RPD can be divided as follows: RPD < 1.0 indicates a very poor model and its use is not recommended; RPD between 1.0 and 1.4 a poor model, where only high and low values are distinguishable; RPD between 1.4 and 1.8 indicates a regular model that allows its use for inferences and correlations; RPD between 1.8 and 2.0 a good model where quantitative predictions are possible; RPD between 2.0 and 2.5 a very good model for quantifications; and RPD > 2.5 an optimum model for quantification.

Results

Sugarcane biomass estimations using a single indicator (univariate regression) showed low efficiency (Table 4). Despite CCCI, all the vegetation indices showed similar ability when estimating biomass. The biometric parameter with the highest R^2 with biomass was HEIGHT, which was comparable to the vegetation indices' performance.

Since biomass prediction using a single variable is limited, multivariable analyses have been proposed in order to improve the models' performance. Approach 2 presented the worst performance in biomass estimation when multivariable models were developed

Table 4 – Metrics obtained by prediction models of biomass with a single indicator.

Variable	R^2	RMSE ¹	RPD ²
HEIGHT	0.52	6.53	1.4
STALKN	0.37	7.49	1.3
SPAD+1	0.10	8.97	1.1
SPAD+3	0.25	8.19	1.2
RED	0.33	7.74	1.2
REDGE	0.00	9.44	1.0
NIR	0.30	7.88	1.2
NDVI	0.49	6.72	1.4
NDRE	0.48	6.79	1.4
CI	0.47	6.89	1.4
MTCI	0.52	6.57	1.4
CCCI	0.01	9.40	1.0
SAVI	0.51	6.61	1.4

¹RMSE = Root mean squared error, in Mg ha⁻¹; ²RPD = Ratio percentage deviation.

(Figure 3C and D), with almost no increase when compared to univariate regressions (NDVI and STALKN - Table 4). Approach 1, which would be the first option when using the proposed method for variable selection, showed a slightly higher result than the other approaches. Approach 3, which maintained complementary agronomic variables, presented an intermediate performance, indicating that the high correlation between the attributes yielded no information gain for modeling.

The degree of importance of each variable for biomass quantification diverged for each approach used (Figures 4A, B and C). However, regardless of the variable selection approach (1 or 2), the individual NIR band and CCCI continued to aggregate information for the models, even though CCCI alone yielded high prediction error (RMSE = 9.4 - Table 4). This happens because the variables that show little meaning when analyzed in isolation can be useful when they are put together in a model (Guyon and Elisseeff, 2003). In addition, chlorophyll readings were not important in any of the approaches adopted.

Discussion

Sensing technology focusing on light reflected by crop canopy is one of the most promising alternatives for estimating crop vigor and, in many cases, it allows for yield site-specific prediction. Such data can be accessed through orbital, aerial and terrestrial platforms. The last alternative in particular, is promising because it can be carried out by agricultural machinery when getting into the field for any kind of management, under any ambient light condition and, in certain situations, it allows for real time interventions. However, we have shown that the information from this type of equipment was not efficient enough to satisfactorily quantify the aboveground biomass of sugarcane (Table

4 - RPD < 1.4, suggesting this is a poor model, where only high and low values were distinguishable). For this purpose, obtaining other plant data has shown to be interesting (greater RPD - all examples on Figure 3A-F). Thus, given a data set with a larger number of variables, we could proceed with classical linear mod-

eling, such as multiple linear regression, or with data mining algorithms, which deal with nonlinear relationships, such as Random Forest. Despite the expected superior performance of Random Forest, we obtained similar results for both modeling algorithms, probably due to the relatively small data set.

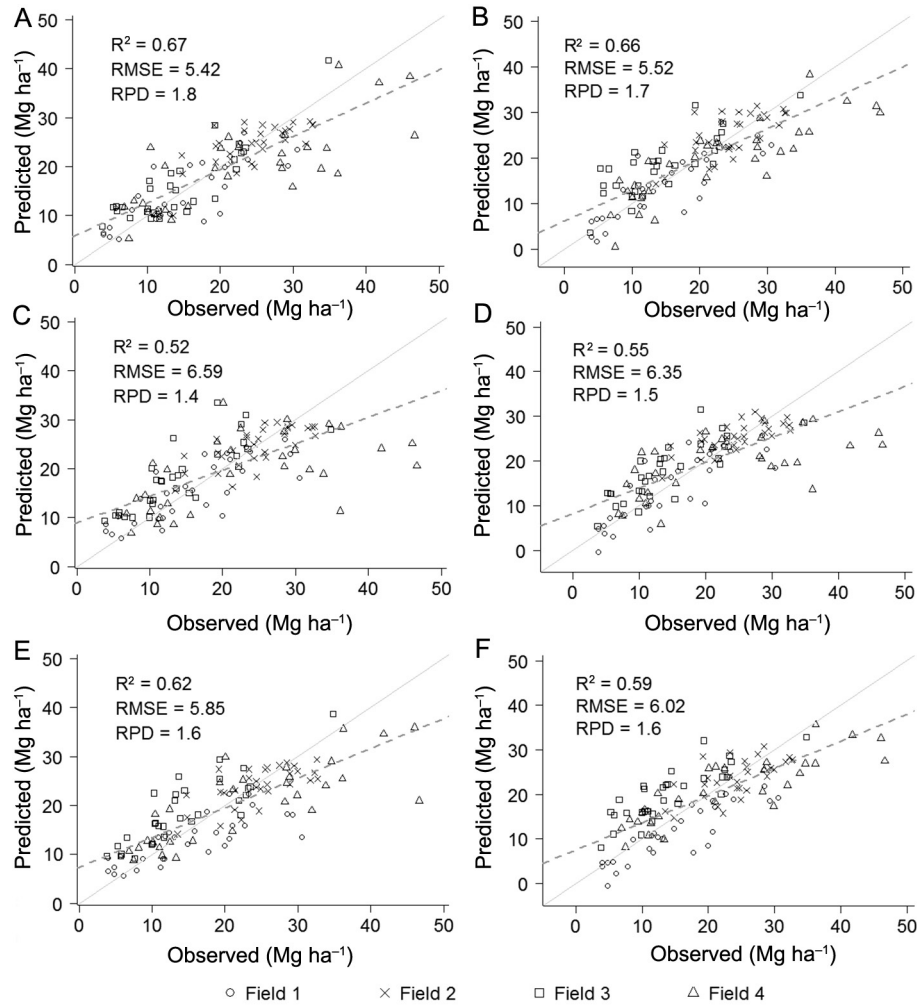


Figure 3 – Predicted and observed biomass correlation of the models estimated by Random Forest for first (A), second (C) and third (E) approach and by Multiple Linear Regression, first (B), second (D) and third (F) approach.

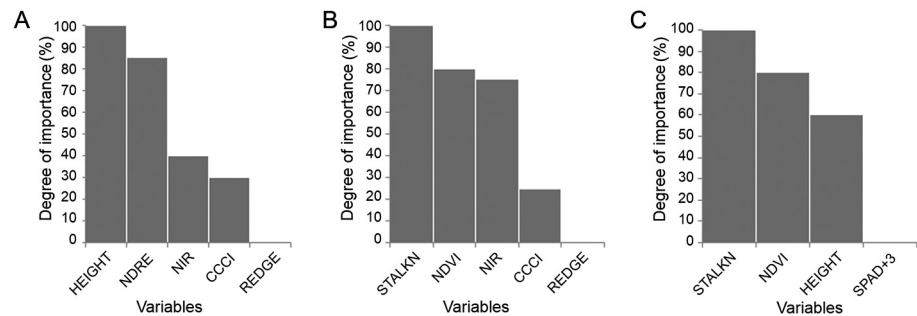


Figure 4 – Degree of importance (normalized metric) of each variable in the Random Forest model for first (A), second (B) and third (C) approaches.

Biomass quantification is inefficient when only one measurement is used (Table 4). This being the case there are limitations as regards reflectance measurements of the sugarcane canopy. The CropCircle sensor shows a nadir (top-down) view of the designated rows, which collected reflectance measurements from the top of the canopy only, impairing its sensitivity to biomass variability. Such limitation was due to the complexity of the sugarcane biomass, which comprised several tillers, each with several layers of green leaves, possibly dried leaves on the stalk inferior part, and the stalk itself, which may have had a different thickness and density.

In this way, it is possible to improve the performance of prediction models through integration with other data sources (as the examples on Figure 3A-F), a concept that has been increasing acceptance in precision agriculture, known as "data fusion". However, the challenge was to develop a sensing technology for such application due to sugarcane being planted mostly throughout the billets, which contributes to a non-uniform stand, as well as presenting differences in the number of tillers between plants (ratoons). The difficulty is even greater when compared to a non-tillering crop (i.e., maize), since a stalk (tiller) may be positioned behind another and not be sensed.

In spite of this, there are tools available which will sense plant height, and can be adapted to sugarcane. This is true in the case of ranging sensors which can be simple ultrasonic sensors for dynamic laser scanning (Light Detection and Ranging - LiDAR technology). This indicated that, to date, Approach 1, as well as presenting the best performance, demonstrated better practical application.

We observed similar performances for the two methods of prediction model construction (Figures 3A, C and E compared to Figures 3B, D and F), despite the expected superiority of the machine learning algorithm. This was probably due to the fact that the dataset was relatively small and did not, therefore, result in very complex relationships between the variables, since this would have been a great gain in machine learning algorithms, as it is more difficult to deal with linear regression. Only in Approach 3 did we see a slightly superior performance by Random Forest, perhaps because it integrated less related measurements to each other which actually add information to the models.

However, biomass quantification needs to be used with caution when the absolute value is important for a particular sugarcane management strategy. Although the use of biometric variables increased the predictive models' capacity, they were all classified as regular performance according to the RPD classification (Viscarra Rossel et al., 2006). Therefore, the models would be recommended only for inferences and correlations, since the errors associated with the models are in the order of 5 to 6 Mg ha⁻¹.

Conclusion

Even though our study underpinned the idea that sugarcane sensing is a complex and problematic issue, we identified that vegetation indices, mainly the NDRE, associated with the number and height of stalks, may assist in sugarcane biomass prediction in the early stages. However, prediction models need to be used with caution when it is critical to quantify the absolute value of biomass.

Biomass prediction through the Random Forest algorithm shows a similar performance to multiple linear regression when used with a small dataset and variables with moderate correlation between each other. Therefore, due to the complexity of using this type of algorithm, linear regression may be preferred by less experienced users.

Acknowledgments

We would like to thank professor José Paulo Molin from the Universidade de São Paulo (ESALQ/USP) for his support in experimental planning and financial support through the PROSENSAP Project - Financiadora de Estudos e Projetos (FINEP) and the students (at that time) Gustavo Portz, Felipe Baricatti Finazzi and Lucas Cortinove for their assistance in field work. This research was carried out under the PhD scholarship of the corresponding author (Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) Project n°2011/08882-7).

Authors' Contributions

Conceptualization: Amaral, L.R., Rocha, M.G. Data acquisition: Amaral, L.R. Data analysis: Rocha, M.G., Oliveira, S.R.M., Barros, F.M.M. Design of Methodology: Amaral, L.R. Writing and editing: Rocha, M.G., Amaral, L.R., Oliveira, S.R.M.

References

- Abdel-Rahman, E.M.; Ahmed, F.B. 2008. The application of remote sensing techniques to sugarcane (*Saccharum spp. hybrid*) production: a review of the literature. *International Journal of Remote Sensing* 29: 3753-3767.
- Amaral, L.R.; Molin, J.P. 2014. The effectiveness of three vegetation indices obtained from a canopy sensor in identifying sugarcane response to nitrogen. *Agronomy Journal* 106: 273-280.
- Amaral, L.R.; Molin, J.P.; Portz, G.; Finazzi, F.B.; Cortinove, L. 2015a. Comparison of crop canopy reflectance sensors used to identify sugarcane biomass and nitrogen status. *Precision Agriculture* 16: 15-28.
- Amaral, L.R.; Molin, J.P.; Schepers, J.S. 2015b. Algorithm for variable-rate nitrogen application in sugarcane based on active crop canopy sensor. *Agronomy Journal* 107: 1513-1523.

- De Benedetto, D.; Castrignanò, A.; Rinaldi, M.; Ruggieri, S.; Santoro, F.; Figorito, B.; Gualano, S.; Diacono, M.; Tamborrino, R. 2013. An approach for delineating homogeneous zones by using multi-sensor data. *Geoderma* 199: 117-127.
- Draper, N.R.; Smith, H. 1998. *Applied Regression Analysis*. 3ed. Wiley, Hoboken, NJ, USA.
- Guyon, I.; Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3: 1157-1182.
- Han, J.; Kamber, M.; Pei, J. 2011. *Data Mining: Concepts and Techniques*. 3ed. Elsevier, Amsterdam, The Netherlands.
- Hastie, T.; Tibshirani, R.; Friedman, J. 2009. *The Elements of Statistical Learning*. 2ed. Springer, Berlin, Germany.
- Lofton, J.; Tubana, B.S.; Kanke, Y.; Teboh, J.; Viator, H.; Dalen, M. 2012. Estimating sugarcane yield potential using an in-season determination of normalized difference vegetative index. *Sensors* 12: 7529-7547.
- Portz, G.; Molin, J.P.; Jasper, J. 2011. Active crop sensor to detect variability of nitrogen supply and biomass on sugarcane fields. *Precision Agriculture* 13: 33-44.
- Raun, W.; Solie, J.; Stone, M.; Martin, K.; Freeman, K.; Mullen, R.; Zhang, H.; Schepers, J.; Johnson, G. 2005. Optical sensor-based algorithm for crop nitrogen fertilization. *Communications in Soil Science and Plant Analysis* 36: 2759-2781.
- Rossi Neto, J.; Souza, Z.M.; Oliveira, S.R.M.; Kölln, O.T.; Ferreira, D.A.; Carvalho, J.L.N.; Braunbeck, O.A.; Franco, H.C.J. 2017. Use of the decision tree technique to estimate sugarcane productivity under edaphoclimatic conditions. *Sugar Tech* 19: 662-668.
- Sandhu, H.S.; Gilbert, R.A.; McCray, J.M.; Perdomo, R.; Eiland, B.; Powell, G.; Montes, G. 2012. Relationships among leaf area index, visual growth rating, and sugarcane yield. *Journal of the American Society of Sugar Cane Technologists* 32: 1-14.
- Tilly, N.; Hoffmeister, D.; Cao, Q.; Lenz-Wiedemann, V.; Miao, Y.; Bareth, G. 2015. Transferability of models for estimating paddy rice biomass from spatial plant height data. *Agriculture* 5: 538-560.
- Viscarra Rossel, R.A.; McGlynn, R.N.; McBratney, A.B. 2006. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma* 137: 70-82.