Accepted version of manuscript *From spreadsheets to sugar content modeling: a data mining approach* This version of the work has been peer-reviewed.
Published version of this work is available at:
https://doi.org/10.1016/j.compag.2016.11.012

Authors' emails were updated and do not necessarily correspond to the ones in the published work.

From spreadsheets to sugar content modeling: a data mining approach

Monique Pires Gravina de Oliveira [a b], Felipe Ferreira Bocca [a c], Luiz Henrique Antunes Rodrigues [a d]

[a] School of Agricultural Engineering, University of Campinas

[b] moniquepgoliveira@gmail.com

[c] -

[d] lique@feagri.unicamp.br - Corresponding author

Abstract

Sugarcane mills need sugar content estimates in advance to establish their commercial strategy. To obtain these estimates, mills rely on historical averages or maturation curves. Crop models have also been developed to provide those estimates. Leveraging mill data about fields and sugar content at harvest, we developed empirical models using different data mining techniques along with the RReliefF algorithm for feature selection. The best model was attained with Random Forest with features selected by RReliefF, having a mean absolute error of 2.02 kg Mg^{-1}. This model outperformed Support Vector Regression and Regression Trees with and without feature selection. Models were also evaluated by the Regression Error Characteristic Curves, which showed that the best model was able to predict 90% of the observations within a precision of 5.40 kg Mg^{-1}.

Keywords

Sugarcane Ripening; Machine Learning; Empirical Modeling; Total Recoverable Sugar; Crop Modeling

1. Introduction

Planning in the sugar industry requires estimates about the amount of sugar that will be produced in the following cycle. This information is then used in forward selling, forward pricing, and managing storage and shipping schedules (Everingham et al., 2007). In Brazil, where harvests mostly occur from April until November, the commercial strategy for the following season starts being established in August of the current season (Bocca et al., 2015). Overestimates could compromise previous selling commitments while underestimates could lead to difficulties in storing and shipping (Everingham et al., 2003). Sugar estimates are also useful for operational level plans, such as prioritizing harvesting areas (Scarpari and Beauclair, 2004).

Two forecasts are required to achieve such estimates: sugarcane (*Saccharum* spp.) fresh mass yield and sugar content in sugarcane stalk (Alvarez et al., 1982; Bocca et al., 2015). The former has been addressed by Everingham et al. (2009, 2016) and by Bocca and Rodrigues (2016). For the latter, industries use either averages from the previous years or variety-specific maturity curves (Scarpari and Beauclair, 2004, 2009). Both approaches, however, do not allow for the inclusion of factors that favor sucrose storage in sugarcane stalks, e.g. weather variability and management practices (van Heerden et al., 2013). Particularly for the case of weather, the increase in weather variability leads to the need of tools to assess the effect of weather uncertainty in production. The urge for climate risk assessment is increasing among companies as an effect of climate change (Surminski, 2013).

To take weather variability and management practices into account, crop models could also be explored. Crop yield models gather knowledge about crop

growth and development and are able to predict its behavior (Boote et al., 1996; Monteith, 1996).

There are mainly two approaches to modeling: to deepen understanding and knowledge of a topic and to make accurate predictions for specific decisions. Frequently, different levels of both can be found in most models (Shmueli, 2010; Singels, 2013). The first approach is seen in models that simulate sugarcane phenological and physiological processes. These models try and describe plants' processes and deepen the understanding of plant physiology and its interactions with the environment (Passioura, 1996; O'Leary, 2000; Singels, 2013). To achieve higher prediction accuracy, aiming at production planning, one could use empirical models, which are independent of the simulations aforementioned.

Empirical models are conceptually simpler models and are based on relationships between crop outputs and its driving factors, e. g. water availability, weather conditions, and agricultural practices (Monteith, 1996; Passioura, 1996; Surendran Nair et al., 2012; Singels, 2013). The relationships explored in these models vary from proxies to direct effects, such as the distance from the lake feature used by Alvarez et al. (1982) and variety, respectively.

Scarpari and Beauclair (2004, 2009) developed empirical models to predict total recoverable sugar by using stepwise regression. In the paper published in 2004, the only variables used by the authors were negative degree-days and available water content during crop development. In 2009, they added another one, concerning photoassimilate production. Despite aiming not to make predictions but to describe the relationship between variables, Lawes et al. (2002) modeled commercial cane sugar by using linear mixed models. Their final model included the year, month of harvest, farm of origin, variety and an interaction

between the month of harvest and year of harvest. More recently, Cardozo et al. (2015) established an exponential relationship between total recoverable sugar and accumulated rainfall in the 120 days before the harvest.

In 1982, Alvarez et al. (1982) had already highlighted not only the vast number of variables that could affect sugarcane yield, but also the complexity of the relations between them. Different approaches have been used to address these issues. Scarpari and Beauclair (2009) generated a set of models: one model was fitted for each combination of variety, number of cuts and type of management zone, for early, mid and late period of harvest during the season. Lawes et al. (2002), in turn, used pairwise combinations of variables, while Cardozo et al. (2015) selected one variable most correlated to sugar content to be included in their three models, for each ripening pattern.

These examples draw attention to the limitations of the methods being used to model sugar content: they either assume linearity, do not extensively account for interactions or both. Also, they should not be directly used for non-normal data with auto-correlated features, which underlines the need for other techniques, such as those highlighted by Breiman (2001), which he called algorithmic models, referring to the models obtained by data mining or machine learning techniques.

Data mining techniques have been long applied in agriculture-related problems, e.g. prediction of wine-fermentation results, evaluation of imperfections in fruits, both with images and X-ray, classification of sounds from pigs and birds, meat analysis and the use of energy in agriculture (Mucherino et al., 2009). The successful application of these techniques is due to their capacity to deal with the previously mentioned aspects of agricultural data.

One further reason to use these other techniques is the availability of data. Lawes et al. (2002) stated that for the Australian production context, some sugar mills collect block-productivity data such as cane yield and commercial cane sugar from every block or paddock harvested during the season, as well as information on the block size, the cane variety, the time of harvest and how many ratoons the cane has. Data collection for Brazilian context is not only similar but also enhanced by the fact that the mill is either owner or responsible for the production (Bocca et al., 2015) and therefore has additional information regarding soil analysis and agricultural practices.

Furthermore, the use of data mining techniques allows for more accurate models since they can identify new and unknown patterns in large datasets (Witten et al., 2011). An attempt in this direction has already been made by Everingham and Sexton (2011), although still with a limited number of variables. It is possible to achieve better estimates by exploring more variables, and by looking at further available algorithms.

Bocca et al. (2015) suggested the use of yield models associated to climate forecasts and production data in an integrated system in order to obtain yield forecasts. In this study, we present the development of an element of this system: a sugar content model that could be used in conjunction with both weather forecasts and production data. To model sugar content (Total Recoverable Sugar - TRS), we use a commercial sugarcane production database and the data mining framework (feature selection, parameter tuning, modeling and validation in an independent set).

2. Materials and methods

2.1. Dataset

Data used in this study were supplied by Alcídia mill, operated by Odebrecht Agroindustrial, located in the city of Teodoro Sampaio, state of São Paulo (SP), Brazil. The mill annual production area is almost 25 thousand hectares of land and its production reaches 1.6 million tons of sugarcane. Harvests that happened in 2011 and 2012 provided 2,102 observations, with each observation referring to one block in the farms in each year. The 53 variables of the dataset belong to four categories: soil physics and soil chemistry, weather, agricultural practices, and those related to the crop (Table 1).

It is worth noting that some variables were created, particularly regarding the developmental stages of the crop, based on the planting dates. Plant cycle was simplified into four stages: (1) sprouting, (2) tillering, (3) growth and (4) maturity. With this approach, we could group weather and phenological information, providing estimates of the weather in each of the plant's stage, rather than averages for the whole cycle.

Variables that delve too much into the cycle, i.e. that are intrinsically related to harvest, such as the occurrence of pests, that is only verified by harvesting time and cannot be predicted in advance, were removed. The remaining variables are either defined in the beginning of the cycle, as is the case for fertilization, or refer to the weather and can be estimated by weather forecasts.

Two scenarios were modeled: (a) using all available features, and (b) performing feature selection and using only the selected features in the modeling process. Feature selection will be further explained in section 2.2.2.

2.2. Model development

2.2.1. Algorithms deployed

In data mining, the prediction of a continuous variable, such as Total Recoverable Sugar, is known as a regression problem. In this paper, three techniques were used to tackle this problem: Support Vector Regression (SVR), Random Forests (RF) and Regression Trees (RT). Statistical software R, version 3.1.1 (R Core Team, 2015) was used in the modeling process with packages *e1071* (Meyer et al., 2014), *randomForest* (Liaw and Wiener, 2002) and *rpart* (Therneau et al., 2014).

2.2.2. Feature selection

Feature selection was only performed on the part of the dataset reserved for training, with 1,402 observations. The algorithm that was chosen to perform feature selection was RReliefF (Robnik-Šikonja and Kononenko, 2003) as it is able to estimate the quality of attributes in problems with strong dependencies between attributes. Since the dataset is comprised of weather and edaphic features, this has turned out to be an important characteristic. The parameters number of neighbors and number of iterations were kept as suggested in Robnik-Šikonja and Kononenko (2003) and Robnik-Sikonja and Kononenko (1997): 10 and 100, respectively. Importance values provided by the algorithm were averaged after 10 repetitions. The RReliefF algorithm implemented in the CORElearn package was used (Robnik-Šikonja et al., 2015).

Robnik-Šikonja and Kononenko (2003) state that importance value is analogous to the percentage of explained variance if scaled to sum 1. Based on that, the criterion to limit the number of features was to select only the best-

ranked attributes that accounted for 0.9 – 90% – of the RReliefF explained variance.

We chose not to perform the variable importance for Regression Tree and Random Forest so that we could use the same variables, chosen by the criteria proposed by RReliefF, in all models.

2.2.3. Parameter tuning

To achieve better results, parameters were tuned using a two-stage grid search. The second stage used smaller step sizes for the grid search in the region close to the best result found in the first stage. Different parameters were searched for the different feature sets used. Details of the ranges and step size for each parameter can be found in the Appendix. Optimal values obtained can be seen in Table 2. Tuning and modeling processes used only data available in the training set, in the two situations modeled, after the groups of features were defined.

2.2.4. Validation

Models were trained in a dataset constituted by two-thirds of the records and validation was achieved through testing in the remaining third (700 observations), still unseen by the model. This strategy is also known as hold-out validation.

Three metrics were computed to report model quality: root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination between model output and observed values (R^2). The REC Curve, developed by Bi and Bennet (2003), was also used with this objective. All metrics of evaluation were reported for the hold-out data.

3. Results and discussion

3.1. Feature Selection

As expected, Algorithm RReliefF was able to identify those variables more often associated with sugar accumulation, e.g. the occurrence of precipitation events and low temperatures close to harvesting (van Heerden et al., 2013). The algorithm was also able to distinguish other variables that could bring information to the model, even though they are less obvious, such as Silt in the intermediate layer. RReliefF selected 31 features from the 53 in the original set (Figure 1).

One can see how weather and crop variables predominate, as not only all of the available ones were selected, but also all of the 13 most important belong to one of those two categories (up to Mean Min Temp 1 - Figure 1). Soil variables were already deemed less important and only 7 remained from the original 24. Even though soil variables are usually autocorrelated, which could have lowered their importance (Robnik-Šikonja and Kononenko, 2003), weather variables are as well. Despite that, these variables had the highest values of importance, highlighting how much they influenced sugar production and accumulation processes.

We do not aim to try and explain every feature chosen by the algorithm, but some will draw attention to other important points. In the already mentioned case of Silt, its selection could have been caused by autocorrelation among features. Along with Texture, they are probably proxies for other soil features and, due to the aforementioned behavior of RReliefF, only Silt and Texture showed enough importance.

The use of soil fractions as features for the models requires a brief discussion. Since soil fractions are compositional data, this characteristic must be

addressed for the use in models such as least squares linear regressions in which independence between features is important and compositional data explicitly violate this assumption. Empirical investigations seem to indicate that such transforms are not required (Ranganathan and Borges, 2011), or that they can be alleviated by the use of GLM models, in which one can specify different error distribution and link functions, or by the use of random forests (Lopatin et al., 2016). In a practical sense, tree methods are invariant to monotonic transformations in the data, meaning these transformations would not change the models. For Support Vector Machines/Regression, if transformations do not change the relative distances between samples, relation between inputs and outputs would not change, but some changes in the internal structure could happen.

Reproducing this methodology with another dataset would probably not present the same outcome, i.e. instead of Silt, Clay could have been chosen. Other ways of engineering the data, such as including latent variables, could also have lead to different outcomes. Despite that, the most important variables are expected to be quite similar.

Another noteworthy result in this selection is the low score presented by the Ripeners feature, which may have occurred due to bias introduced in the dataset. In both cases, when ripeners were applied and when they were not, TRS would never present an excessively low value, or else the area would not have been harvested. Since low TRS fields were not harvested, they did not make it into the commercial dataset. This illustrates how biases might often be unaccounted for in the modeling process.

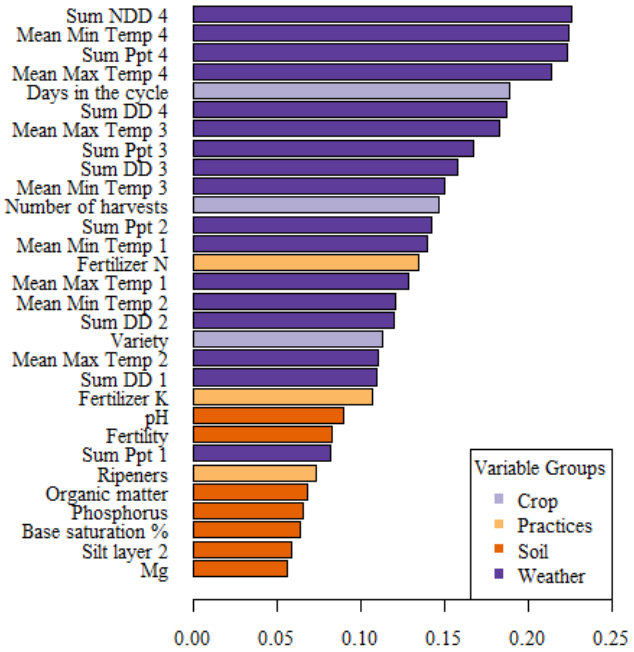

Figure 1. Ranked feature importance of features selected by the RReliefF algorithm. Weather features names follow the pattern of Mean/Sum, variable studied and period for the calculation (1 for sprouting, 2 for tillering, 3 for growth and 4 for maturity), e.g. Sum DD 4 is the sum of degree days in the maturity period and Mean Max Temp 1 is the mean value of the daily maximum air temperature for the sprouting period.

This unexpectedness of some selected features, nevertheless, creates opportunities for investigating other unknown variables that play a role in sucrose accumulation in sugarcane stalks.

3.2. Modeling

On average, developed models achieved mean absolute error values that represent about 2.5% of the smallest value in the test set. This result means that

decision makers are prone to forecast errors that depend almost exclusively on the quality of the weather forecast models available.

Based on MAE or RMSE values, one could affirm the best prediction outcome came from algorithm Random Forest both when feature selection was performed and when it was not (Table 3). However, the REC curve allows for further understanding of the prediction errors (Figure 2). We observe that once the acceptable error is larger than 15 kg Mg^{-1}, all models present similar performance. This means criteria other than the accuracy of the predictions might be used, e.g. processing speed and interpretability. Regression trees often have an advantage of interpretability (Breiman, 2001), but taking into account the complexity parameter achieved through tuning, this is not the case for the models obtained. The trees turned out to have more than 150 nodes.

One should note that not only the Random Forest mean absolute error is 2.02 kg Mg^{-1}, but that 80% of the predicted values have a mean absolute error lower than 3.00 kg Mg^{-1} and that 90% have an error lower than 5.40 kg Mg^{-1}. These limits – if acceptable to decision makers – reinforce the model's utility in the decision support process.

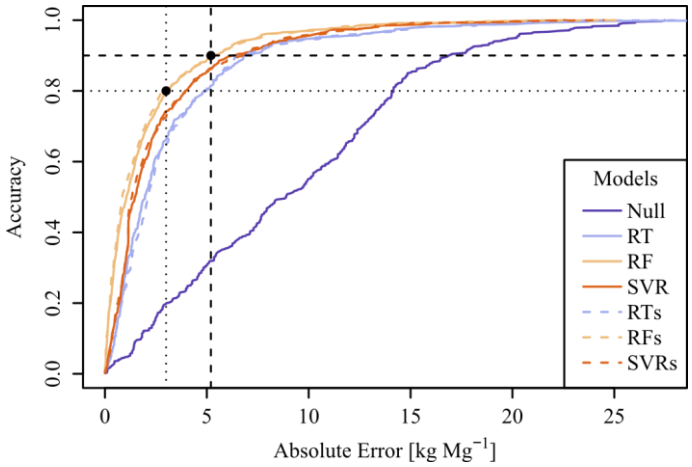

Figure 2. REC Curve. Dotted line indicates an error threshold of 3.00 kg Mg^{-1}. Dashed line indicates an error threshold of 5.40 kg Mg^{-1}. Letter "s" indicates feature selection.

Even though the coefficient of determination between model output and observed values should not be used as a validation metric (Harrison, 1990; Mitchell, 1997), we do so (Figure 3) as this is the only reported metric regarding the validation process in some of previous research papers.

We stress that comparisons of the present models with previous works should consider the differences in the validation strategy. Everingham and Sexton (2011) used a cross-validation approach and obtained R^2 of 0.55, while Scarpari and Beauclair (2004) reported R^2 of the training set of 0.70.

Likewise, we believe our results should not be compared to the ones obtained by Scarpari and Beauclair (2009) and by Cardozo et al. (2015) due to the different strategies of validation. In the former paper, several models were developed and different coefficients are presented for each model. In the latter, measured values of TRS are averages of all of the blocks in each city, weighted by the proportion of each group of cultivars, for each month. The average value is then compared to the predicted TRS for each month of harvest. Due to this

aggregation, observed data variability is reduced, diminishing residuals from comparison between predicted and observed values.

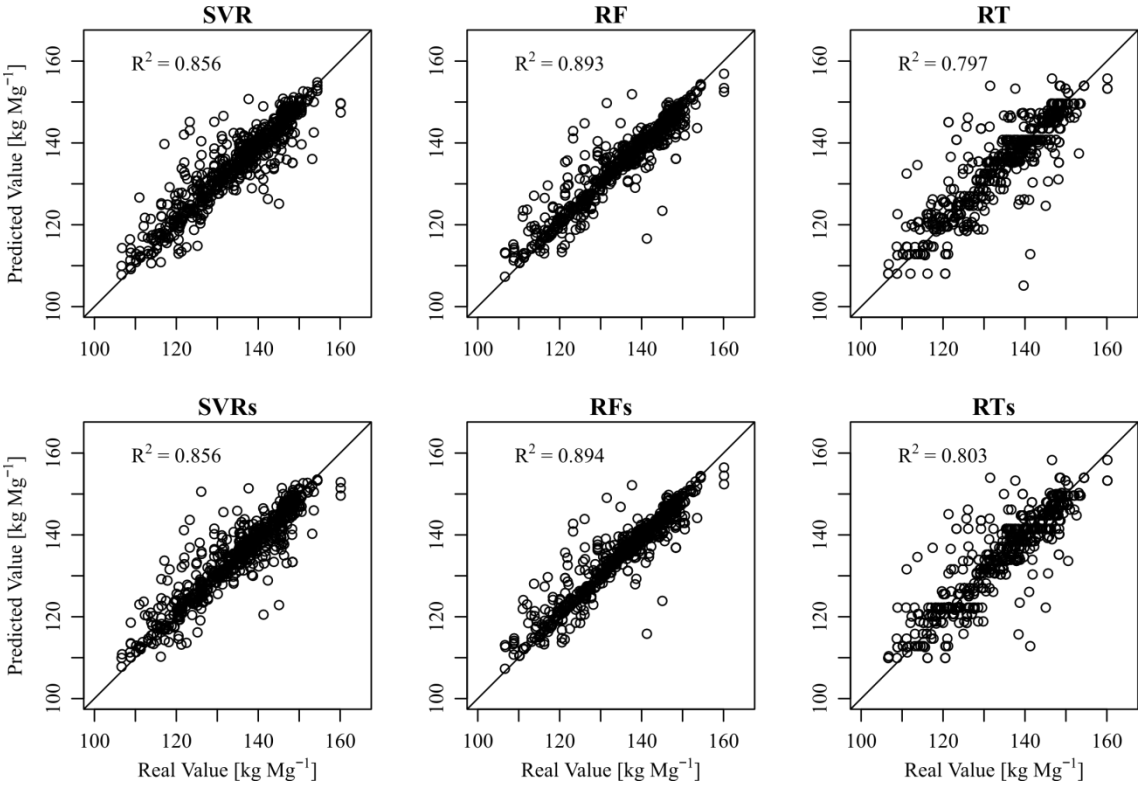

Figure 3. Real values vs. predicted values obtained by the different models. Techniques used were Support Vector Regression (SVR), Random Forests (RF) and Regression Tree (RT). Straight line indicates the 1:1 reference. Letter "s" indicates feature selection.

A data mining approach is different from those used in previous modeling, not only regarding the choice of which variables would be included in the model, but also because it allows for variables that are less generic than Farm and Year. Lawes et al. (2004) pointed out that the Farm and Year variables capture the influence of management and local variations in the environment. In variables such as month or year there is an implicit consideration of past weather, which

happened during the months of growth of the observations in the training set. Breaking these variables into several others allows the algorithms to extract the knowledge that each individually brings to the pool, as well as prevent inclusion in the model of information that would not be useful. Also, when variables are explicitly considered, they allow for including estimates, such as the amount of nitrogen applied for the cycle.

We acknowledge that the modeling errors do not correspond to forecasting errors, since weather data used correspond to past data. These models, when used with forecasting intentions, have to be used in combination with weather forecasting technologies (Everingham et al., 2007). In the system proposed by Bocca et al. (2015), in what concerns production data, some estimates should be included, such as amount of fertilizer applied, ripener application, etc.

Therefore, the highest importance of weather variables in the last period has both a positive and negative impact on the potential use of the models for operational decision-making. Concerning decisions regarding harvesting dates – which can be changed late in the crop cycle – weather forecasts will be needed only a few months in advance, leading to better estimates. On the other hand, decisions that depend on long-range forecasts will have larger errors for the latest months.

As for feature selection, one can note how similar the results were when the technique was performed and when it was not. We expected improvements in the results, since the technique was developed to deal with performance degradation both in speed and in performance accuracy due to high dimensionality (Kira and Rendell, 1992). Even though there was no pronounced improvement, the similar

performance despite the lower number of variables implies that the developed models were more robust.

We underline that there are also other well-known techniques that could have been explored, e.g. neural networks, boosted models or even creating an ensemble of the models. However, due to its iterative nature, data-mining could be indefinitely performed for improved performance (more data, more features, refined tuning, etc.), and we have already shown the better performance of data mining techniques.

Finally, we must add that due to the scope of the data used to fit our models, their use should be limited to the mill that provided the data. At most, they could be used by mills with similar conditions of soil and weather. If the conditions differ, our workflow should be repeated with data from the different mill.

4. Conclusions

Data mining techniques not only showed potential in Total Recoverable Sugar prediction at block level, but also showed an improvement when compared to previous modeling attempts. We attributed this improvement to the inclusion of more detailed data and to the refinement of these techniques. Developed models were allowed to explore relationships other than linear or first degree interactions.

By exploring a feature selection technique, we reinforced the importance of weather variables and we also could detect the introduction of bias in the dataset. If the technique is to be used, said bias should be removed by either not including blocks in which ripeners were applied or by including data before and after ripeners were applied.

We evaluated models according to several metrics, and all of them supported the use of these techniques in decision making in the sugar industry, considering the availability of appropriate weather data. Particularly, the use of the REC curve provided ranges of accuracy that can also be used by decision makers.

5. Acknowledgements

This work was supported by Bioen/Fapesp and Odebrecht Agro-industrial (Process # 12/50049-3 - Link: http://www.bv.fapesp.br/en/auxilios/55622/data-mining-techniques-applied-to-the-analysis-and-prediction-of-sugarcane-yield/); and CAPES (Coordination for the Improvement of Higher Education Personnel).

6. References

Alvarez, J., Crane, D.R., Spreen, T.H., Kidder, G., 1982. A yield prediction model for Florida sugarcane. Agr. Syst. 9, 161–179. doi:10.1016/0308-521X(82)90018-X

Bi, J., Bennet, K.P., 2003. Regression error characteristic curves, in: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). Washington DC, pp. 43–50.

Bocca, F.F., Rodrigues, L.H.A., 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Comput. Electron. Agric. 128, 67–76. doi:10.1016/j.compag.2016.08.015

Bocca, F.F., Rodrigues, L.H.A., Arraes, N.A.M., 2015. When do I want to know and why? Different demands on sugarcane yield predictions. Agr. Syst. 135, 48–56. doi:10.1016/j.agsy.2014.11.008

Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of

400 crop models. Agron. J. 88, 704–716.
401 doi:10.2134/agronj1996.00021962008800050005x

402 Breiman, L., 2001. Statistical Modeling: The Two Cultures. Stat. Sci. 16, 199–
403 231. doi:10.1214/ss/1009213726

404 Cardozo, N.P., Sentelhas, P.C., Panosso, A.R., Palhares, A.L., Ide, B.Y., 2015.
405 Modeling sugarcane ripening as a function of accumulated rainfall in
406 Southern Brazil. Int. J. Biometeorol. doi:10.1007/s00484-015-0998-6

407 Demattê, J.L.I., Demattê, J.A.M., 2009. Ambientes de produção como estratégia
408 de manejo na cultura da cana-de-açúcar. Informaçoes Agronômicas 10–18.

409 Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate
410 prediction of sugarcane yield using a random forest algorithm. Agron.
411 Sustain. Dev. 36, 27. doi:10.1007/s13593-016-0364-z

412 Everingham, Y.L., Inman-Bamber, N.G., Thorburn, P.J., McNeill, T.J., 2007. A
413 Bayesian modelling approach for long lead sugarcane yield forecasts for the
414 Australian sugar industry. Aust. J. Agric. Res. 58, 87–94.
415 doi:10.1071/AR05443

416 Everingham, Y.L., Muchow, R.C., Stone, R.C., Coomans, D.H., 2003. Using
417 southern oscillation index phases to forecast sugarcane yields: A case study
418 for northeastern Australia. Int. J. Climatol. 23, 1211–1218.
419 doi:10.1002/joc.920

420 Everingham, Y.L., Sexton, J., 2011. An introduction to multivariate adaptive
421 regression splines for the cane industry, in: Proceedings of the 2011
422 Conference of the Australian Society of Sugar Cane Technologists.

423 Everingham, Y.L., Smyth, C.W., Inman-Bamber, N.G., 2009. Ensemble data
424 mining approaches to forecast regional sugarcane crop production. Agric.
425 For. Meteorol. 149, 689–696. doi:10.1016/j.agrformet.2008.10.018

426 Harrison, S.R., 1990. Regression of a model on real-system output: An invalid test
427 of model validity. Agr. Syst. 34, 183–190. doi:10.1016/0308-

521X(90)90083-3

Kira, K., Rendell, L., 1992. A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 249–256.

Lawes, R.A., McDonald, L.M., Wegener, M.K., Basford, K.E., Lawn, R.J., 2002. Factors affecting cane yield and commercial cane sugar in the Tully district. Aust. J. Exp. Agric. 42, 473–480. doi:10.1071/EA01020

Lawes, R.A., Wegener, M.K., Basford, K.E., Lawn, R.J., 2004. The evaluation of the spatial and temporal stability of sugarcane farm performance based on yield and commercial cane sugar. Aust. J. Agric. Res. 55, 335–344. doi:10.1071/AR03169

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Lopatin, J., Dolos, K., Hernández, H.J., Galleguillos, M., Fassnacht, F.E., 2016. Comparing Generalized Linear Models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. Remote Sens. Environ. 173, 200–210. doi:10.1016/j.rse.2015.11.029

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.

Mitchell, P.L., 1997. Misuse of regression for empirical validation of models. Agr. Syst. 54, 313–326. doi:10.1016/S0308-521X(96)00077-7

Monteith, J.L., 1996. The quest for balance in crop modeling. Agron. J. 88, 695–697. doi:10.2134/agronj1996.00021962008800050003x

Mucherino, A., Papajorgji, P.J., Pardalos, P.M., 2009. Data Mining in Agriculture, Springer Optimization and Its Applications. Springer New York, New York, NY. doi:10.1007/978-0-387-88615-2

O'Leary, G.J., 2000. A review of three sugarcane simulation models with respect

455 to their prediction of sucrose yield. Field Crop. Res. 68, 97–111.
456 doi:10.1016/S0378-4290(00)00112-X

457 Passioura, J.B., 1996. Simulation models: Science, snake oil, education, or
458 engineering? Agron. J. 88, 690–694.
459 doi:10.2134/agronj1996.00021962008800050002x

460 R Core Team, 2015. R: A Language and Environment for Statistical Computing.

461 Ranganathan, Y., Borges, R.M., 2011. To transform or not to transform. Plant
462 Signal. Behav. 6, 113–116. doi:10.4161/psb.6.1.14191

463 Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of
464 ReliefF and RReliefF. Mach. Learn. 53, 23–69.
465 doi:10.1023/A:1025667309714

466 Robnik-Šikonja, M., Kononenko, I., Konokenko, I., 1997. An adaptation of Relief
467 for attribute estimation in regression, in: Proceedings of the Fourteenth
468 International Conference on Machine Learning - ICML '97. Morgan
469 Kaufmann Publishers Inc., San Francisco, CA, pp. 296–304.

470 Robnik-Šikonja, M., Savicky, P., Alao, J.A., 2015. CORElearn: Classification,
471 Regression and Feature Evaluation.

472 Scarpari, M.S., Beauclair, E.G.F. de, 2009. Physiological model to estimate the
473 maturity of sugarcane. Sci. Agric. 66, 622–628. doi:10.1590/S0103-
474 90162009000500006

475 Scarpari, M.S., Beauclair, E.G.F. de, 2004. Sugarcane maturity estimation through
476 edaphic-climatic parameters. Sci. Agric. 61, 486–491. doi:10.1590/S0103-
477 90162004000500004

478 Shmueli, G., 2010. To Explain or to Predict? Stat. Sci. 25, 289–310.
479 doi:10.1214/10-STS330

480 Singels, A., 2013. Crop Models, in: Sugarcane: Physiology, Biochemistry, and
481 Functional Biology. John Wiley & Sons Ltd, pp. 541–577.

doi:10.1002/9781118771280.ch20

Surendran Nair, S., Kang, S., Zhang, X., Miguez, F.E., Izaurralde, R.C., Post, W.M., Dietze, M.C., Lynd, L.R., Wullschleger, S.D., 2012. Bioenergy crop models: Descriptions, data requirements, and future challenges. GCB Bioenergy 4, 620–633. doi:10.1111/j.1757-1707.2012.01166.x

Surminski, S., 2013. Private-sector adaptation to climate risk. Nat. Clim. Chang. 3, 943–945. doi:10.1038/nclimate2040

Therneau, T., Atkinson, B., Ripley, B., 2014. rpart: Recursive Partitioning and Regression Trees.

van Heerden, P.D.R., Eggleston, G., Donaldson, R.A., 2013. Ripening and Postharvest Deterioration, in: Sugarcane: Physiology, Biochemistry, and Functional Biology. John Wiley & Sons Ltd, pp. 55–84. doi:10.1002/9781118771280.ch4

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, Data Mining: Practical Machine Learning Tools and Techniques. Elsevier. doi:10.1016/B978-0-12-374856-0.00001-8

Appendix

Table 4. Ranges explored in the parameter tuning process.

Algorithms	Parameters	First tuning	Second tuning
SVR	Gamma	From 2^{-9} to 2^{1}. Step: 2^{n}	20 points linearly spaced from 2^{-6} to 2^{-3}.
	Cost	From 2^{0} to 2^{10}. Step: 2^{n}	10 points linearly spaced from 2 to 8.
RF	Percentages of attributes in the split	15%, 22%, 33%, 44%, 55%, 66%, 77%	10 points linearly spaced from 33% to 55%.
	Minimum size of terminal nodes	1,2,5,10,20	1,2
	Number of trees to grow	100,250,500,750, 1000, 1500	20 points linearly spaced from 250 to 1500.
RT	Minimum number of objects in a node for a split	2,5,10,25,50	-
	Complexity parameter	0.001, 0.005,0.01,0.05,0.1, 0.25	-

FIGURE CAPTIONS

Figure 1. Ranked feature importance of features selected by the RReliefF algorithm. Weather features names follow the pattern of Mean/Sum, variable studied and period for the calculation (1 for sprouting, 2 for tillering, 3 for growth and 4 for maturity), e.g. Sum DD 4 is the sum of degree days in the maturity period and Mean Max Temp 1 is the mean value of the daily maximum air temperature for the sprouting period.

Figure 2. REC Curve. Dotted line indicates an error threshold of 3.00 kg Mg^{-1}. Dashed line indicates an error threshold of 5.40 kg Mg^{-1}. Letter "s" indicates feature selection.

Figure 3. Real values vs. predicted values obtained by the different models. Techniques used were Support Vector Regression (SVR), Random Forests (RF) and Regression Tree (RT). Straight line indicates the 1:1 reference. Letter "s" indicates feature selection.

Table 1. List of variables used for modeling in the full dataset.

	Variable	Details [Unit]
Soil	Sand, Clay, and Silt	Numeric variables. Each percentage is available in three depths: 0-0.25 m, 0.25-0.50 m and 0.80-1.00 m [%].
	Texture	Categorical variable with seven levels. The percentage of clay in the soil is the main criteria used to define the levels.[(a)]
	Fertility	Categorical variable with seven levels. Levels are determined based on soil percent base saturation, soil aluminum saturation, and soil pH. [(a)]
	Soil Density	Numeric variable [g cm^{-3}].
	Chemical properties	Numeric variables. Extractable aluminum [$mmol_c$ dm^3], extractable calcium [$mmol_c$ dm^{-3}], extractable potassium [$mmol_c$ dm^{-3}], extractable magnesium [$mmol_c$ dm^{-3}], soil pH, phosphorus concentration [mg dm^{-3}], organic matter [g dm^{-3}].
	Derived chemical properties	Numeric variables. Cation-exchange capacity [$mmol_c$ dm^{-3}], exchangeable acidity [$mmol_c$ dm^{-3}], percent base saturation [%], aluminum saturation [%], and sum of bases [$mmol_c$ dm^{-3}].
Weather	Sum of degree days (Sum DD), Accumulated	Numeric variables. Each one is available for the four stages of plant's

	precipitation (Sum Ppt), Mean maximum air temperature (Mean Max Temp), Mean minimum air temperature (Mean Min Temp)	development: (1) sprouting, (2) tillering, (3) growth and (4) maturity.
	Sum of Negative Degree Days (Sum NDD 4)	Numeric variable. Only available in the maturity stage.[(b)]
Agricultural practices	Ripeners	Categorical variable referring to whether ripeners were applied or not
	Vinasse	Numerical variable referring to the irrigation depth using vinasse [mm]
	Cake Filter	Numerical variable referring to the amount of cake filter applied [Mg ha^{-1}]
	Fertilization rates of Nitrogen, Phosphorus, Potassium and Molybdenum.	Rates were calculated according to applied bulk rate and formula percentages. Particularly for Potassium, rate could also have been determined by using its concentration in vinasse and how much vinasse was applied. [kg ha^{-1}]
	Source of Potassium	Categorical variable referring to whether K source is either mineral fertilizer or vinasse.
Crop related	Variety	Categorical variable
	Days in the cycle	Numeric variable referring to the interval from planting or previous harvest to the

	following harvest
Number of harvests	Categorical variable with labels 1 to 6. Harvests that happened within 12, 15 and 18 months after planting or previous harvest are hardcoded as 1.
Total Recoverable Sugar (TRS)	Numeric variable for the mass of sugar in a metric ton of sugarcane. Expressed in [kg Mg^{-1}].

(a) As proposed by Dematt\u00ea and Dematt\u00ea (2009). (b) Negative degree days as described by Scarpari and Beauclair (2004).

Table 2. Parameters defined by tuning for the full dataset and for the dataset with selected features.

Algorithms	Parameters	Full dataset	Dataset with feature selection
SVR	Gamma	0.02138158	0.06743421
	Cost	6	4
RF	Percentages of attributes in the split	43% (23 attributes)	43% (13 attributes)
	Minimum size of terminal nodes	1	1
	Number of trees to grow	579	382
RT	Minimum number of objects in a node for a split	2	2
	Complexity parameter	0.001	0.001

Table 3. Mean Absolute Errors and Root Mean Squared Errors of data mining techniques in the considered scenarios (kg Mg^{-1}).

	Features used	Regression Tree	Random Forest	SVR	Null model*
MAE	All features	3.23	2.08	2.66	9.26
	Selected features	3.27	2.02	2.64	9.26
RMSE	All features	5.06	3.62	4.18	11.05
	Selected features	4.98	3.60	4.19	11.05

*Predicted values correspond to the mean TRS value in the training set.