Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

From spreadsheets to sugar content modeling: A data mining approach

Monique Pires Gravina de Oliveira, Felipe Ferreira Bocca, Luiz Henrique Antunes Rodrigues *

School of Agricultural Engineering, University of Campinas, Av. Cândido Rondon, 501-Barão Geraldo, 13083-875 Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form 12 November 2016
Accepted 17 November 2016

Keywords:
Sugarcane ripening
Machine learning
Empirical modeling
Total recoverable sugar
Crop modeling

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}.

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,

* Corresponding author.
E-mail addresses: monique.oliveira@feagri.unicamp.br (M.P.G. de Oliveira), felipe.bocca@feagri.unicamp.br (F.F. Bocca), lique@feagri.unicamp.br (L.H.A. Rodrigues).

http://dx.doi.org/10.1016/j.compag.2016.11.012
0168-1699/© 2016 Elsevier B.V. All rights reserved.

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 autocorrelated 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 2102 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 1402 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-Šikonja et al. (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