|Application of data mining to forest operations planning

Rossit, Daniel Alejandro[1]; Olivera [2], Alejandro; Viana Céspedes, Víctor[2]; Broz, Diego[3]

[1] Departamento de Ingeniería, Universidad Nacional del Sur, CONICET
Av. Alem 1253, Bahía Blanca (B8000CPB), Argentina.
daniel.rossit@uns.edu.ar
[2] Universidad de la República
Ruta 5, km 386,5, Tacuarembó (C.P. 45000), Uruguay.
alejandro.olivera@cut.edu.uy, vviana@cci.edu.uy
[3] UNaM CONICET Facultad de Ciencias Forestales
Bertoni 124, Eldorado (N3382GDD), Misiones, Argentina.
diegoricardobroz@gmail.com

Abstract

In Uruguay, mechanized forestry harvesting for industrial purposes is carried out using modern equipment. They are capable of record a wealth of information that can be exploited in the decision making process and improve operations. Some approaches from data mining field, as decision trees, are an alternative to analyze large volumes of data and determine incidence factors. In this work, it was proposed to analyze how different variables of the forest harvest (DBH [1], species, shift and operator) affect the productivity of a forest harvester. Data were collected automatically by a forest harvester working on plantations of *Eucalyptus spp*. in Uruguay. The results show that DBH is the most influential factor in productivity.

1 Introduction

Forest planning is a highly complex decision-making problem involving various factors: ecological, productive and economic systems ([1], [2], [3]). A large extent of this complexity is due to the duration of biological processes involved, such as tree growth, since the length of rotations can reach 25 years [4]. This makes the planning of harvest operations complex and affects the economic performance of companies. Forest harvesting is key factor in commercial forest plantations because of its high impact on production costs, quality and value recovery of forest products (mainly wood) and, also, on their potential environmental impact. In this sense, estimating the productivity (measured in m^3h^{-1}) of these activities is a central issue for planning harvesting operations efficiently. Therefore, a precise estimation of harvester productivity will contribute to improve the supply chain of forest products (from the field to the industry).

Forest harvesting operations in Uruguay use modern machines equipped with automatic data collection technology. This fact makes available a large amount of harvest data that can be processed using data mining techniques for later use in harvest planning and forest management. Olivera et al [5] studied the productivity of harvesting operations in *Eucalyptus spp*. Plantations in Uruguay using data automatically collected by a harvester. With this data, the authors performed a regression analysis to study the effect of five variables on the machine productivity. The variables that significantly affected productivity in order of importance were: Diameter at Breast Height (DBH) of the trees, operator, and work shift (day and night). However, the regression analysis method only allows comparing the dependent variable productivity with a single independent variable at a time, something that limits a more integrative view of the system. In this paper we propose to revisit this problem, using a data mining approach, specifically, classification or decision trees (DT). This methodology will allow a more accurate description of the dependent variable productivity by analyzing its dependence on a set of variables at a time, instead of a single variable. According to Ahlemeyer-Stubbe and Coleman [6], DTs are popular and reliable methods for developing prediction and classification models. For the best of our knowledge, there is no evidence of the application of this technique in forest harvesting planning, although it is a very versatile technique for exploratory data analysis. The objective of this work is to apply this technique using a data set collected automatically by a forest harvester to evaluate the productivity of the operation.

1 DBH: diameter at breast height.

2 Metodologhy

DT methodology is widely disseminated in the field of data mining. It consists of generating a prediction model of a dependent variable as a function of a set of independent variables. The generated model is a tree, in which each branch describes rules in terms of the independent variables that allow to predict categories of the dependent variable with a good level of approximation. This model is based on the exploration of a set of observations. In this paper, we use the classification tool of the SPSS IBM software, and CHAID (Chi-square Automatic Interaction Detector) as analysis procedure.

Our case study comprises a data set of 4805 records of processed trees, obtained from data collected by a forest harvester working on plantations of *Eucalyptus spp*. The machine registers a time stamp when it each tree is fall. We calculated the cycle time of each processed tree determining the difference between two consecutive records as explained in [5]. In addition to the time stamp, the machine also records for each tree: harvested volume (m^3) and DBH. Complementary information was included as variables that can affect productivity: species, shift (day / night) and operator. Productivity was the dependent variable, which was converted into categorical variable, where each category indicates a range of productivity. Next, the decision tree method formulates rules to predict the occurrence of each productivity category. In this work, we propose a gross categorization of productivity to be able to present the methodology. These categories are too broad for a real practical purpose, but to discretize in lower range would imply a larger number of categories, which would turn this work little illustrative and a cumbersome example. The categories adopted are 4 and were named by their upper bound, the first is "<12" considering productivities below 12 m^3h^{-1}, the second: "<= 26" for productivities between 12 and 26 m^3h^{-1}, the third: "<= 40" for productivities between 26 and 40 m^3h^{-1}, and the last one, "> 40" for productivities that exceed 40 m^3h^{-1}.

Once the decision tree model is validated, it can be used to predict the forest harvester productivity in similar situations. This prediction is valuable for planning forest operations and, consequently, for forest product supply chain management.

3 Results and Discussion

Figure 1 illustrates the results obtained by DT models. In the DT model, the dependent variable is Productivity The node 0, shows the observations obtained for each category. Then, the first branch uses the variable "DBH" to classify the observations, generating 8 nodes. From nodes 1 to 8 in each node there is a dominant category of productivity, and, each dominant category has a percentage greater than the percentage at node 0 (greater purity). For example, in the node 1, where DBH is below or equal to 122 mm, the most likely category of productivity is "<12" with almost 80% probabilities. At the next level, nodes 6 and 7 are branched and the variable "Operator" is used to classify the observations. At this level "Operator" values are indicated for each sheet, where the purity of the nodes is improved, as in node 9, which allows to improve the productivity prediction "> 40", from 41.6% to 51.2%. Then, the prediction rules are read from the leaf node to the root node. For example, the rule for node 9 (a leaf node) is: if "Operator" is operator 1, and the DBH is between 204 and 235 mm the productivity will probably be "> 40". In contrast, for node 10, the difference would be "Operator" is operator 2, then the productivity is likely to be "<= 40".

The "Shift" variable did not have a significant effect, so productivity does not significantly depend on the shift or, at least, does to a lesser extent than "DBH" and "Operator". This agrees with [5] and [7], which validates the proposed method, at least, in a gross manner for this brief instance.

The DT method allowed establishing differences in productivities between operators (1 and 2) for DBH classes between 204 and 274 mm for *E, dunnii*. This result allows a more accurate quantification of productivity differences between operators than the regression analysis presented in [5] and [7]. In the mentioned work, the significant difference between operators is established, but the methodology (regression analysis) did not allow to detail for which diameter classes that difference was significant. This level of detail if possible by applying the DT technique. In practice, this level of detail allows to allocate operators to work in different strata of a plantation of this species of *Eucalyptus* according to the expected productivity. In strata with diameters smaller than 204mm, either operator will have similar productivity, while above this diameter it is expected Operator 1 to be more productive. This tree (Figure 1) is a concise example of the approach proposed in this paper, which shows the great potential of such approaches for forest operations planning. For having a detail of the precision of the model statistical assessments are done. The results of these assessments are included in Table

1 as Confusion matrix table. For further research, we will study the influence of other variables on productivity and extend the analysis to other dependent variables and larger datasets.

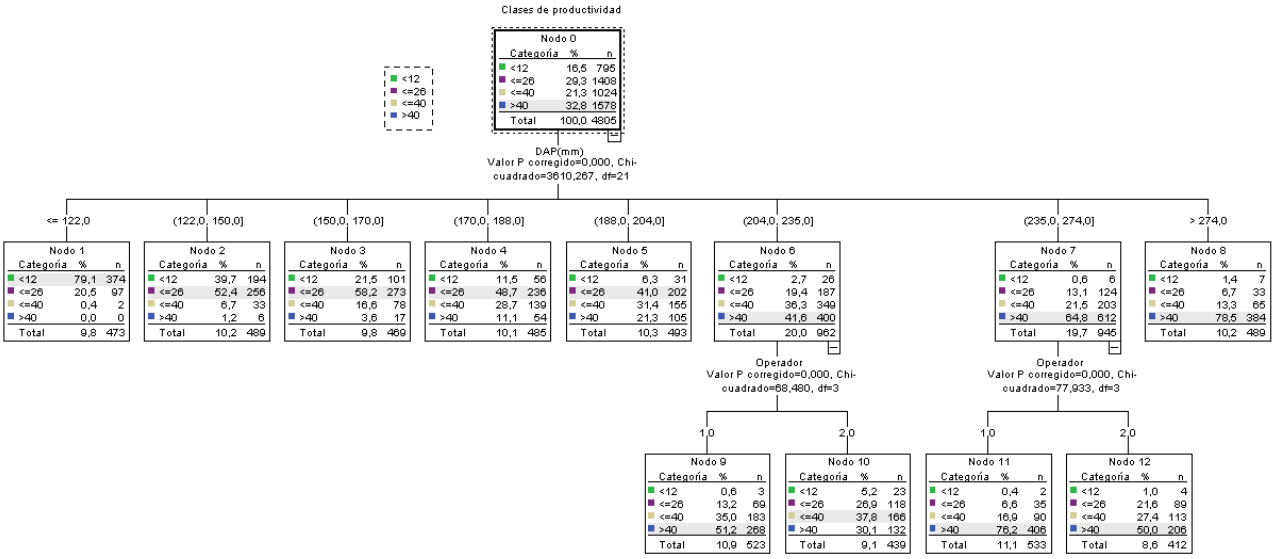

Figure 1. Decision tree model to predict the kind of productivity a harvester.

Observed	***Forecasted***				
	<12	***<=26***	***<=40***	***>40***	***Correct percentage***
< 12	374	382	23	16	47%
<=26	97	967	118	226	68.7%
<=40	2	405	166	451	16.2%
>40	0	182	132	1264	80.1%
Global percentage	9.8%	40.3%	9.1%	40.7%	57.7%

Table 1. Confusion matrix of the tree model.

Referencias

[1] Weintraub, A., y Romero, C. (2006). Operations research models and the management of agricultural and forestry resources: a review and comparison. Interfaces, 36(5), 446-457.

[2] Milanesi, G. S., Broz, D., Tohmé, F., & Rossit, D. (2014). Strategic analysis of forest investments using real option: The fuzzy pay-off model (FPOM). *Fuzzy Economic Review*, *19*(1), 33.

[3] Broz, D., Durand, G., Rossit, D., Tohmé, F., y Frutos, M. (2016). Strategic planning in a forest supply chain: a multigoal and multiproduct approach. *Canadian Journal of Forest Research*, *47*(999), 297-307.

[4] Broz, Diego Ricardo. "Técnicas de simulación y optimización aplicadas a la planificación forestal." (2015). Ediuns, Bahía Blanca, Argentina.

[5] Olivera, A., Visser, R., Acuna, M., & Morgenroth, J. (2015). Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a *Eucalyptus spp*. harvesting operation in Uruguay. *International Journal of Forest Engineering*, *27*(1), 15-28.

[6] Ahlemeyer-Stubbe, A., Coleman, S. *A practical guide to data mining for business and industry*, John Wiley & Sons, 2014.

[7] Olivera Farias, Ernesto Alejandro. "Exploring opportunities for the integration of GNSS with forest harvester data to improve forest management." (2016).