Revista Pădurilor 138(4) (2023) 001-066

Revista Pădurilor

available online at: www.revistapadurilor.com

PERFORMANCE OF K-MEANS CLUSTERING ALGORITHM IN FINDING SUITABLE GROUPS: A CASE STUDY ON OPERATIONAL PERFORMANCE DATA OF A HARVESTER

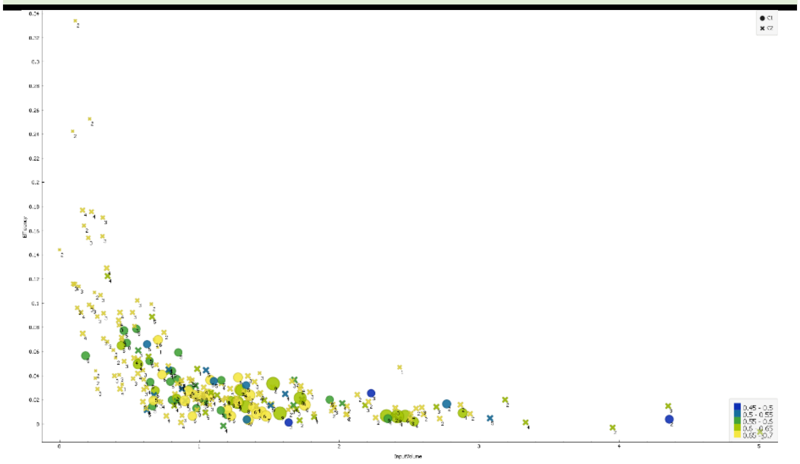
Stelian Alexandru Borz[a,*]

[a]**Department of Forest Engineering, Forest Management Planning and Terrestrial Measurements, Faculty of Silviculture and Forest Engineering, Transilvania University of Braşov, Şirul Beethoven 1, Braşov 500123, Romania, stelian.borz@unitbv.ro (S.A.B.).**

HIGHLIGHTS

- k-means clustering technique was used to extract meaningful categories of data in terms of work cycle time, efficiency and productivity.
- Single-feature clustering solutions provide a good differentiation in the data range of features used.
- Clustering solutions may be useful assuming that mean values of the target variables are used for differentiation in performance.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Manuscript received: 01 September 2023
Received in revised form: 24 November 2023
Accepted: 01 December 2023
Page count: 20 pages.

Article type: Research Article

Editor: Stelian Alexandru Borz

Keywords:
Timber harvesting
Efficiency
Differentiation
k-means clustering
Piece-rate systems

ABSTRACT

Piece-rate systems are typically used in timber harvesting to reflect the variation in performance based on increment or decrement in the values of inputs or other operational factors. Heterogenous data which typically comes from time studies is sometimes difficult to categorize based on the observed values. In this study, the k-means clustering method is used to find meaningful categories in the data sourced by observations on a harvester which processed delimbed pieces of various lengths and input volumes in a number of 1 to 9 logs. A dataset containing more than 230 observations was used to cluster the data on work cycle time, efficiency and productivity based on the input volume, piece length and number of recovered logs. The results indicate that single feature-based clustering solutions provide the best differentiation in the range of that feature but not in the range of the target data used. However, the performance metrics such as the efficiency and productivity were well separated by their mean values after clustering, making the method used valuable for finding useful information for piece-rate setting systems.

* Corresponding author. Tel.: +40-742-042-455.
E-mail address: stelian.borz@unitbv.ro

1. INTRODUCTION

Increasing the economic efficiency in forest operations has compelled many contractors in purchasing advanced equipment able to completely mechanize the operations in timber harvesting. In many European countries, harvesters are being increasingly used to fell and process trees [1] and an important market has been established lately for such equipment also in Romania. Although such machines are used typically to fell and process the trees, for reasons such as a higher productivity and safety, as well as for getting a better machine utilization rate, they are being also used to process trees, tree lengths or long logs at the roadside [2, 3]. When working with logs or tree lengths, such machines are used to crosscut them into the intended lengths, which typically qualifies the resulted pieces as final assortments. However, the logs or tree lengths may come in various sizes, whereas the size is a factor affecting the performance of processing [4-6]. In addition, the decisions taken on the size of the recovered logs will also affect the time consumption and productivity of processing operations. As a consequence, all of these will affect the economic performance.

It is common to use time studies to evaluate the performance of timber harvesting operations in terms of efficiency and productivity [6-9], and there are many international studies which focused on evaluating the performance, developing models of time consumption and productivity for harvesters and processors, comparing their performance in different operational conditions as well as with other means used in tree felling and processing [2, 3, 5, 6, 10-17]. In turn, the assessment of productive performance is important for cost estimation [18, 19]. Previous studies included a certain variability in tree size or in other factors, that enabled the development of mathematical models to relate the time consumption and/or productivity to relevant operational factors. Sometimes, however, there might be a high variability in the values of factors used to predict the performance of mechanized operations. When working in mechanized processing tasks at the forest road, the tree lengths or logs may come in various lengths and diameters, while the decision on bucking the logs at a given length may not be related to these factors. These will limit the ability of a model to explain the variability in time consumption or productivity. In other cases, one may choose to establish homogeneous data groups and to use the average values as indicators in performance, an approach that is used in some piece-rate setting systems; they take as explanatory variables categories of factors such as the average tree size, extraction distance and species group, and provide figures on expected efficiency and productivity [20].

Setting up categories in factors and performance metrics typically follows the logic of having a higher performance as the volume of the input work object increases, that is, the higher the input volume the higher the productivity. In heterogeneous decision-making conditions, however, one may decide to recover differentially the logs in a way that is not necessarily a function of the size of the input work object. This will affect the processing performance and will complicate further the establishment of homogeneous categories.

This study uses an unsupervised clustering technique to check whether the data can be meaningfully categorized based on a set of feature variables such as the input volume, piece length and number of recovered logs, and a target variable such as the time consumption, efficiency or productivity. Taking as input the database from [3], the study was designed to iteratively run a k-means clustering algorithm over the data to identify meaningful clusters in the attempt to answer to the question on whether the method used is suitable in finding well differentiated groups in feature and target data so as to minimize the overlapping in their data range. This was complemented by the characterization of data which was done by descriptive statistics.

2. MATERIALS AND METHODS

2.1. Data sourcing and specification

A dataset documented at elemental level was used in this study (**Table 1**), reflecting the variability in performance of timber processing at the forest road by a single grip harvester.

Table 1. Description of the data used in the study.

Parameter (abbreviation)	Measurement unit	Main descriptive statistics
Time consumption		
Swinging to grab (t_{sg})	Seconds	n = 232, min. value = 4, max value = 119, mean±standard deviation value = 28.3±17.94
Grabbing (t_g)	Seconds	n = 232, min. value = 1, max value = 46, mean±standard deviation value = 6.8±6.14
Swinging to process (t_{sp})	Seconds	n = 231, min. value = 2, max value = 227, mean±standard deviation value = 32.8±25.35
Processing (t_p)	Seconds	n = 229, min. value = 1, max value = 66, mean±standard deviation value = 6.7±5.73
Arranging & piling (t_{ap})	Seconds	n = 231, min. value = 2, max value = 98, mean±standard deviation value = 20.8±14.92
Work cycle time (T)	Hours	n = 232, min. value = 0.006, max value = 0.119, mean±standard deviation value = 0.026±0.014
Explanatory variables		
Input volume (v)	m^3	n = 232, min. value = 0.1, max value = 5.0, mean±standard deviation value = 1.2±0.84
Piece length (l)	m	n = 232, min. value = 4.5, max value = 26.0, mean±standard deviation value = 13.0±4.20
Number of recovered logs (n)		n = 232, min. value = 1, max value = 9, mean±standard deviation value = 3.9±1.57
Performance metrics		
Efficiency (E)	$h \times m^{-3}$	n = 232, min. value = 0.002, max value = 0.336, mean±standard deviation value = 0.040±0.045
Productivity (P)	$m^3 \times h^{-1}$	n = 232, min. value = 2.979, max value = 418.605, mean±standard deviation value = 57.762±57.754

The dataset provided the background data for the modeling study of [3] and it featured detailed observations of the main work elements of processing, as well as on explanatory variables such as the input volume of the pieces, their length and number of recovered logs per piece. Based on the input volume and the time spent in processing tasks, the database contained also estimates in efficiency and productivity for each processed piece.

As shown in **Table 1**, a work cycle was divided into five work elements, namely (i) swinging to grab, (ii) grabbing, (iii) swinging to process, (iv) processing, and (v) arranging and piling. Swinging to grab consisted of moving the machine's boom to the piece to be processed, grabbing consisted of securing a given piece into the processor head, swinging to process consisted of moving the machine's boom to a location or between the locations at which crosscutting was done, processing consisted of the effective crosscutting, and arranging and piling consisted of supplementary movements to arrange and pile the processed logs. Efficiency and productivity were computed based on the input volume of a given piece and the cycle time spent to process that piece. For analysis, the following variables were used: work cycle time (hereafter *T*), input volume (hereafter *v*), piece length (hereafter *l*), number of recovered logs (hereafter *n*), efficiency (hereafter *E*), and productivity (hereafter *P*).

2.2. Data clustering

k-means clustering is a method that was designed to partition a set of observations (n) into a number of clusters (k), in which each observation belongs to the cluster with the nearest mean that serves as a prototype of the cluster. The standard algorithm of the method was first proposed by Lloyd [21] and it became lately known as the Lloyd-Forgy algorithm. Conceptually, the k-means clustering algorithm belongs to the group of unsupervised learning and clustering techniques, serving to finding patterns by grouping the data based on selected features and a target variable. The method minimizes within cluster variance and it supports random or more advanced initializations; in this study, the Orange Visual Programming software [22] was used to run the clustering tasks taking as an option a random initialization of the clusters. The supplementary options used in clustering were the following: the solution enabled the formation 2 to 10 clusters; ten re-runs were selected for clustering, and the maximum number of iterations was set to 10000.

Clustering scenarios (**Table 2**) were designed to successively take as target variables the work cycle time (*T*), efficiency (*E*), and productivity (*P*). In the used dataset, these were calculated for each entry based on the **Equations 1-3**.

$$T\ [\mathrm{h}] = (t_{sg}\ [\mathrm{s}] + t_{g}\ [\mathrm{s}] + t_{sp}\ [\mathrm{s}] + t_{p}\ [\mathrm{s}] + t_{ap}\ [\mathrm{s}])\ /\ 3600 \qquad (1)$$

$$E\ [\mathrm{h} \times \mathrm{m}^{-3}] = T\ [\mathrm{h}]\ /\ v\ [\mathrm{m}^{3}] \qquad (2)$$

$$P\ [\mathrm{m}^{3} \times \mathrm{h}^{-1}] = v\ [\mathrm{m}^{3}]\ /\ T\ [\mathrm{h}] \qquad (2)$$

Note: the description of the members shown in **Equations 1-3** are given in **Table 1**.

Table 2. Scenarios used for data clustering.

Scenario	Features	Target	Description
SvT	*v*	*T*	k-means clustering taking as a feature the input volume and as a target the cycle time
SlT	*l*	*T*	k-means clustering taking as a feature the piece length and as a target the cycle time
SnT	*n*	*T*	k-means clustering taking as a feature the number of recovered logs and as a target the cycle time
SvlT	*v, l*	*T*	k-means clustering taking as features the input volume and piece length and as a target the cycle time
SvnT	*v, n*	*T*	k-means clustering taking as features the input volume and number of recovered logs and as a target the cycle time
SlnT	*l, n*	*T*	k-means clustering taking as features the piece length and number of recovered logs and as a target the cycle time
SvlnT	*v, l, n*	*T*	k-means clustering taking as features the input volume, piece length and number of recovered logs and as a target the cycle time
SvE	*v*	*E*	k-means clustering taking as a feature the input volume and as a target the efficiency
SlE	*l*	*E*	k-means clustering taking as a feature the piece length and as a target the efficiency
SnE	*n*	*E*	k-means clustering taking as a feature the number of recovered logs and as a target the efficiency
SvlE	*v, l*	*E*	k-means clustering taking as features the input volume and piece length and as a target the efficiency
SvnE	*v, n*	*E*	k-means clustering taking as features the input volume and number of recovered logs and as a target the efficiency
SlnE	*l, n*	*E*	k-means clustering taking as features the piece length and number of recovered logs and as a target the efficiency
SvlnE	*v, l, n*	*E*	k-means clustering taking as features the input volume, piece length and number of recovered logs and as a target the efficiency
SvP	*v*	*P*	k-means clustering taking as a feature the input volume and as a target the productivity
SlP	*l*	*P*	k-means clustering taking as a feature the piece length and as a target the productivity
SnP	*n*	*P*	k-means clustering taking as a feature the number of recovered logs and as a target the productivity
SvlP	*v, l*	*P*	k-means clustering taking as features the input volume and piece length and as a target the productivity
SvnP	*v, n*	*P*	k-means clustering taking as features the input volume and number of recovered logs and as a target the productivity
SlnP	*l, n*	*P*	k-means clustering taking as features the piece length and number of recovered logs and as a target the productivity
SvlnP	*v, l, n*	*P*	k-means clustering taking as features the input volume, piece length and number of recovered logs and as a target the productivity

Variables used as features were the input volume (*v*), number of recovered logs (*n*) and the piece length (*l*). The number of clustering scenarios was designed by considering any possible combination of feature variables, resulting in a total number of 21 scenarios named by the target and the feature variables used (**Table 2**). For each scenario, the quality of clustering, as well as the number of clusters retained as final was evaluated based on the silhouette score. The silhouette score is a metric used to evaluate the goodness of a clustering technique [23], and it can take values from -1 to 1. A value of 1 indicates a solution with clearly distinguishable clusters, a value of 0 indicates that clusters are indifferentiable, and a value of -1 indicates that the clusters were assigned in a wrong way.

2.3. Statistical analysis

As a first step of data visualization, the initial data was plotted in bi-variate plots by taking as a dependent variable the efficiency (E, h × m^{-3}) and as independent variables the input volume (v) and the piece length (l), respectively. The developed plots included the categorization of data as a function of number of recovered logs (n). Two plots were developed this way, with the aim of showing the main effects of independent variables over the efficiency. Then, for each clustering solution retained as final for a scenario, the data was taken from a data table widget (**Figure 1**) and transferred into a Microsoft Excel ® sheet. Here, the main descriptive statistics such as the minimum, maximum, mean, and standard deviation values were computed for each cluster and for each feature and target variable at the scenario level. Based on the standard deviation and mean values, the coefficients of variation were computed for each cluster from a given scenario.

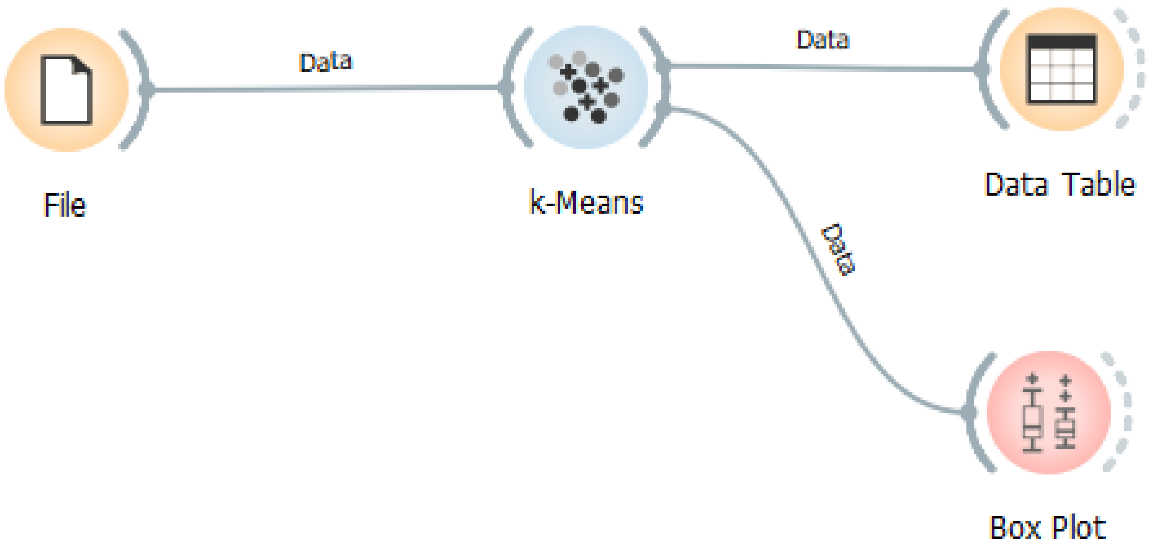


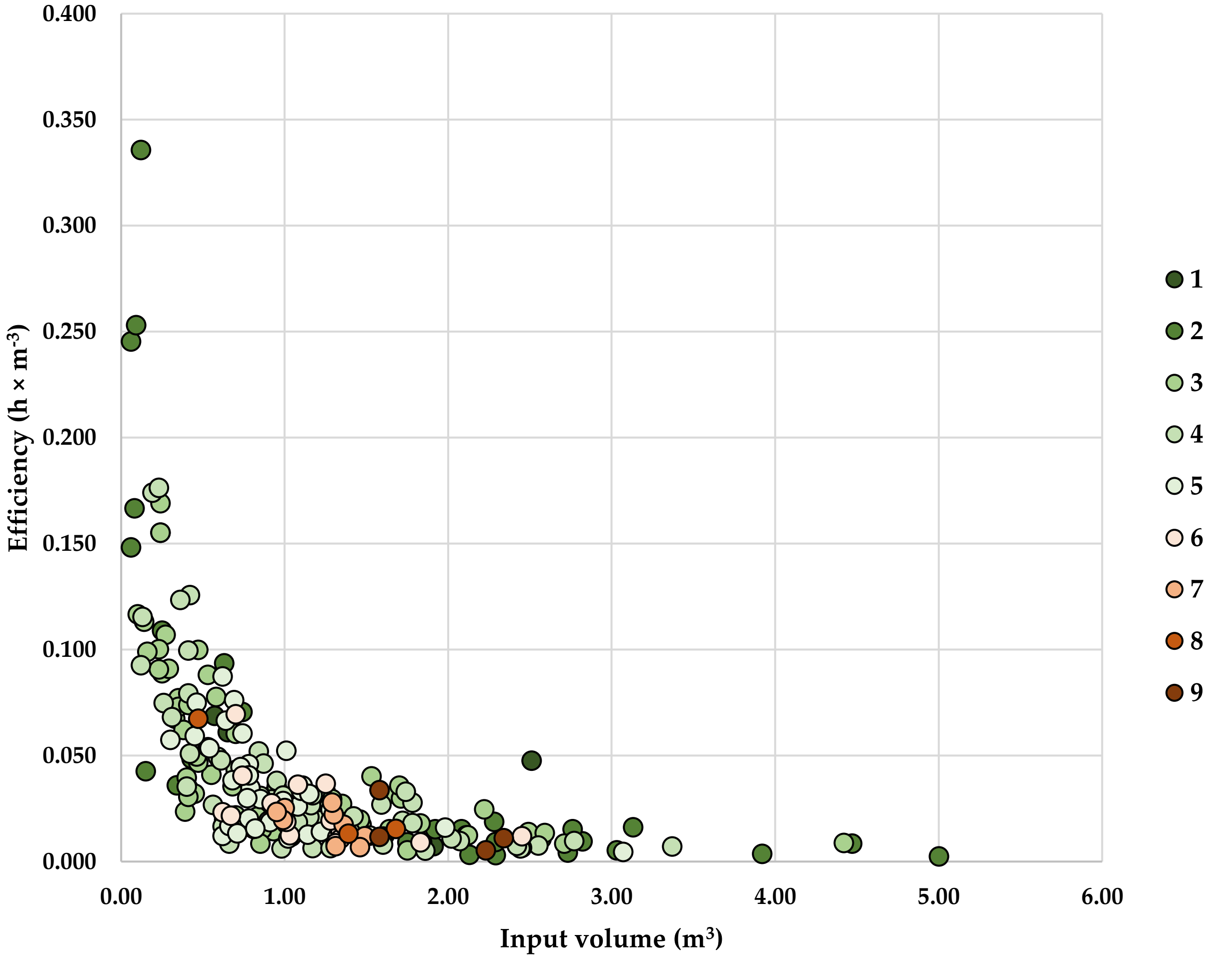
Figure 1. The workflow used in Orange Visual Programming software for clustering, data visualization and extraction.

A Box Plot widget was used to visualize the grouping of feature data for each clustering solution. The same widget was used to extract figures showing the main descriptive statistics of feature data, including its dispersion, which were used to document further the best clustering solutions in terms of target variables. The best clustering solutions for target variables were considered to be those providing a good segmentation of data in its range, which was evaluated based on the way in which the data ranges overlapped.

3. RESULTS

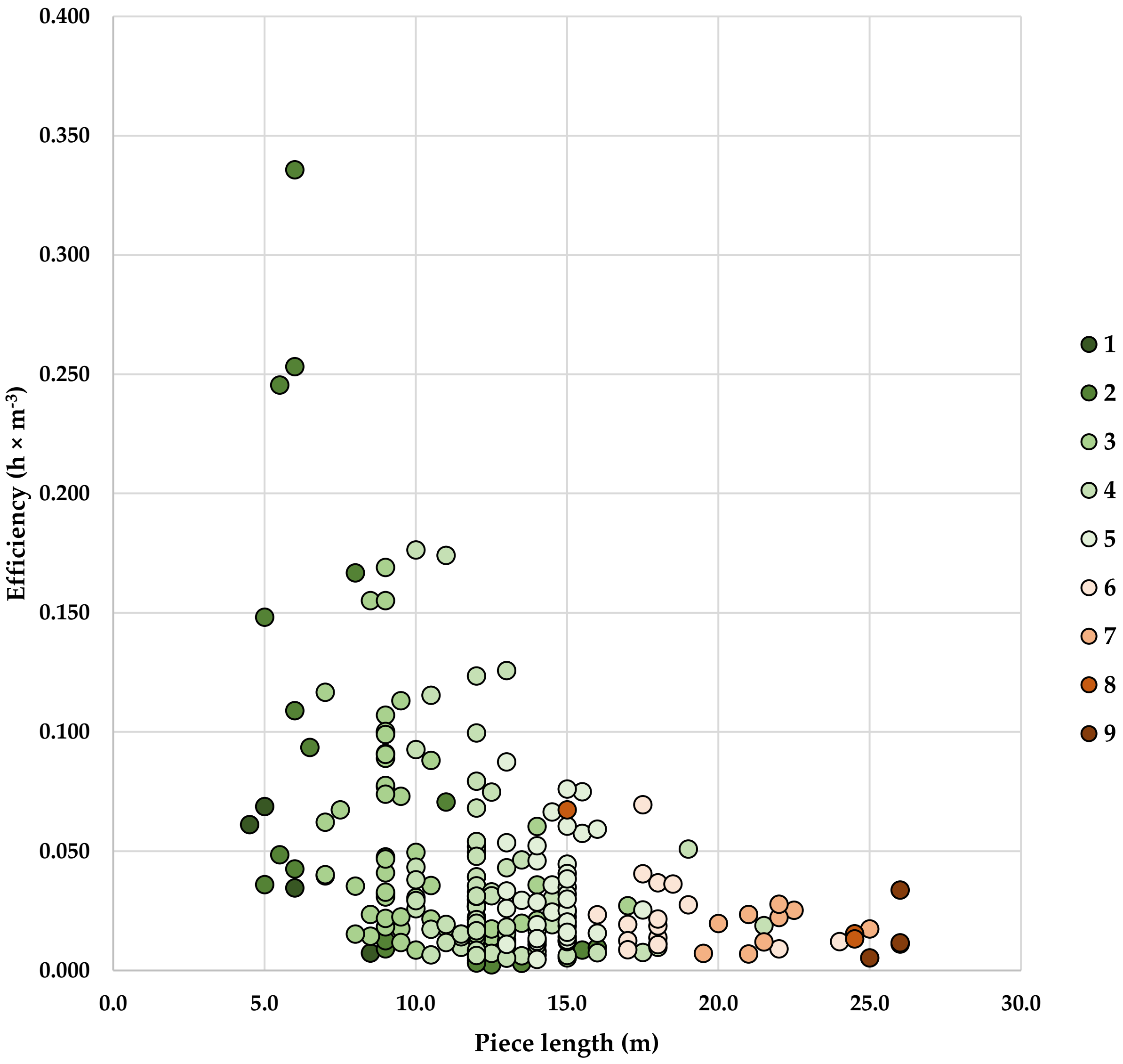
3.1. Visualization of initial data

Figure 2 shows the variation in efficiency as a function of input volume (panel **a**) and piece length (panel **b**), by considering also the number of recovered logs (*n*). The general trend was that to have a higher efficiency as the input volume (panel **a**) and piece length (panel **b**) increased (please note that lower figures of efficiency indicate a higher efficiency).



a

Figure 2. Variation in efficiency as a function of input volume (a) and piece length (b) by taking into consideration the number of recovered logs (*n*, right side of each figure panel).



b

Figure 2, continued. Variation in efficiency as a function of input volume (a) and piece length (b) by taking into consideration the number of recovered logs (*n*, right side of each figure panel).

As the **Figure 2** shows, however, the number of recovered logs was not necessarily well correlated with the input volume; for instance, there were instances in which a single log was recovered for input volumes of less than 0.5 m^3, as well as for pieces having more than 2 m^3. Piece length was more correlated with the number of recovered logs, but also in this case there were overlaps in data. For instance, a number of three logs was recovered from pieces having lengths of about 10 m, but also from some having lengths of 17 m or more. The same was found when recovering a single log, which was typical for pieces of about 5 m in length, but occurred also for pieces or 12 to 15 m.

3.2. Data clustering solutions

Tables 3-5 show the main descriptive statistics of the clustering solutions such as the minimum, maximum, mean, and standard deviation values along with the coefficients of variation.

Table 3. Clustering solutions for scenarios taking the work cycle time as a target variable.

Scenario (Silhouette score)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
SvT	0.007	0.006	-	-	-	-	-	-	-	-
(0.629)	0.119	0.061	-	-	-	-	-	-	-	-
	0.029	**0.026**	-	-	-	-	-	-	-	-
	0.018	0.012	-	-	-	-	-	-	-	-
	62.90	47.63	-	-	-	-	-	-	-	-
SlT	0.009	0.007	0.006	0.010	0.006	0.008	0.009	0.012	0.007	-
(0.679)	0.119	0.052	0.061	0.036	0.057	0.055	0.049	0.053	0.061	-
	0.026	**0.025**	**0.026**	**0.022**	**0.029**	**0.027**	**0.024**	**0.026**	**0.026**	-
	0.019	0.013	0.017	0.008	0.012	0.012	0.011	0.012	0.015	-
	70.86	49.87	65.50	33.73	43.26	45.05	47.22	48.31	57.19	-
SnT	0.007	0.006	0.014	0.006	0.009	0.011	0.008	0.012	0.018	-
(1.000)	0.061	0.057	0.119	0.059	0.036	0.049	0.054	0.053	0.032	-
	0.028	**0.025**	**0.045**	**0.025**	**0.021**	**0.024**	**0.027**	**0.027**	**0.025**	-
	0.013	0.012	0.043	0.014	0.008	0.012	0.012	0.018	0.007	-
	45.37	47.59	95.78	56.82	39.63	50.26	45.58	67.05	26.44	-
SvlT	0.009	0.006	0.006	-	-	-	-	-	-	-
(0.536)	0.053	0.061	0.119	-	-	-	-	-	-	-
	0.024	**0.027**	**0.026**	-	-	-	-	-	-	-
	0.010	0.013	0.017	-	-	-	-	-	-	-
	44.00	47.29	64.96	-	-	-	-	-	-	-
SvnT	0.009	0.008	0.009	0.007	0.006	0.006	0.007	0.012	0.009	-
(0.533)	0.053	0.054	0.061	0.119	0.059	0.053	0.047	0.039	0.057	-
	0.024	**0.026**	**0.032**	**0.029**	**0.026**	**0.025**	**0.025**	**0.026**	**0.026**	-
	0.011	0.012	0.015	0.023	0.016	0.011	0.011	0.015	0.013	-
	46.12	46.93	46.05	77.46	61.19	46.12	43.89	56.48	51.56	-
SlnT	0.007	0.006	0.006	0.007	0.008	0.008	0.012	0.007	0.011	0.009
(0.528)	0.061	0.061	0.057	0.119	0.053	0.054	0.053	0.053	0.049	0.036
	0.032	**0.026**	**0.026**	**0.027**	**0.026**	**0.026**	**0.026**	**0.029**	**0.024**	**0.021**
	0.018	0.017	0.012	0.017	0.011	0.014	0.012	0.012	0.011	0.008
	56.52	64.08	48.10	65.17	44.24	54.49	48.31	41.664	46.29	38.91
SvlnT	0.008	0.006	-	-	-	-	-	-	-	-
(0.500)	0.055	0.119	-	-	-	-	-	-	-	-
	0.025	**0.027**	-	-	-	-	-	-	-	-
	0.011	0.015	-	-	-	-	-	-	-	-
	45.46	54.64	-	-	-	-	-	-	-	-

Note: from top to bottom, in each cell of the table are the minimum, maximum, mean, standard deviation values and the coefficient of variation of the target variable; mean values are given in bold, and C1 to C10 stand for the clustering solutions.

Table 4. Clustering solutions for scenarios taking the efficiency as a target variable.

Scenario (Silhouette score)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
SvE	0.002	0.006	-	-	-	-	-	-	-	-
(0.629)	0.048	0.336	-	-	-	-	-	-	-	-
	0.014	**0.049**	-	-	-	-	-	-	-	-
	0.009	0.048	-	-	-	-	-	-	-	-
	68.65	98.12	-	-	-	-	-	-	-	-
SlE	0.007	0.007	0.035	0.007	0.002	0.005	0.007	0.005	0.003	-
(0.679)	0.169	0.176	0.336	0.028	0.123	0.076	0.069	0.034	0.126	-
	0.064	**0.054**	**0.108**	**0.018**	**0.027**	**0.031**	**0.024**	**0.015**	**0.027**	-
	0.048	0.051	0.092	0.007	0.023	0.021	0.017	0.008	0.026	-
	75.52	95.51	84.941	40.36	88.08	67.16	69.82	56.10	96.60	-
SnE	0.008	0.008	0.035	0.003	0.002	0.005	0.007	0.005	0.003	-
(1.000)	0.169	0.123	0.336	0.336	0.123	0.076	0.069	0.034	0.126	-
	0.044	**0.038**	**0.108**	**0.053**	**0.027**	**0.031**	**0.025**	**0.017**	**0.027**	-
	0.040	0.027	0.092	0.099	0.023	0.020	0.017	0.010	0.026	-
	92.31	71.15	84.94	187.62	86.73	66.32	68.72	58.94	96.60	-
SvlE	0.005	0.002	0.007	-	-	-	-	-	-	-
(0.536)	0.069	0.174	0.336	-	-	-	-	-	-	-
	0.021	**0.029**	**0.072**	-	-	-	-	-	-	-
	0.014	0.027	0.064	-	-	-	-	-	-	-
	66.20	91.49	88.94	-	-	-	-	-	-	-
SvnE	0.023	0.002	0.007	0.003	0.253	0.017	0.008	0.010	-	-
(0.533)	0.034	0.174	0.336	0.015	0.336	0.123	0.169	0.033	-	-
	0.028	**0.030**	**0.067**	**0.010**	**0.294**	**0.046**	**0.063**	**0.017**	-	-
	0.005	0.027	0.062	0.005	0.058	0.029	0.051	0.009	-	-
	18.38	89.58	93.14	50.10	19.83	62.42	80.79	53.12	-	-
SlnE	0.003	0.035	0.006	0.007	0.006	0.005	0.005	0.002	0.008	0.007
(0.528)	0.060	0.336	0.174	0.176	0.076	0.126	0.034	0.071	0.069	0.028
	0.022	**0.106**	**0.041**	**0.061**	**0.035**	**0.030**	**0.015**	**0.017**	**0.025**	**0.017**
	0.017	0.090	0.038	0.048	0.021	0.028	0.008	0.012	0.017	0.008
	80.13	84.64	92.79	78.60	91.92	93.77	56.10	74.61	69.49	45.37
SvlnE	0.005	0.002	-	-	-	-	-	-	-	-
(0.500)	0.076	0.336	-	-	-	-	-	-	-	-
	0.026	**0.046**	-	-	-	-	-	-	-	-
	0.018	0.051	-	-	-	-	-	-	-	-
	71.88	110.37	-	-	-	-	-	-	-	-

Note: from top to bottom, in each cell of the table are the minimum, maximum, mean, standard deviation values and the coefficient of variation of the target variable; mean values are given in bold, and C1 to C10 stand for the clustering solutions.

By the highest silhouette score, the final solutions contained between two and ten clusters. In addition, the number of clusters was the same when considering a given set of features used, irrespective of the used target variable. The silhouette scores ranged from 0.500 to 1.000, with the latter characterizing the solutions clustered by the number of the recovered logs, which was a discrete variable. None of the solutions provided a clear separation of the values of target variables by considering the range of variation.

In terms of mean values (**Tables 3-5**, **Figure 3**), the differentiation of data was improved, particularly when considering the performance metrics such as the efficiency or productivity as target variables. In turn, the use of work cycle time provided clusters with average values of the target variable which were close together (**Figure 3a**).

Table 5. Clustering solutions for scenarios taking the productivity as a target variable.

Scenario (Silhouette score)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
SvP	21.014	2.979	-	-	-	-	-	-	-	-
(0.629)	418.605	160.364	-	-	-	-	-	-	-	-
	111.725	**38.938**	-	-	-	-	-	-	-	-
	78.882	31.319	-	-	-	-	-	-	-	-
	70.60	80.43	-	-	-	-	-	-	-	-
SlP	5.918	5.671	2.979	36.000	8.100	13.143	14.400	29.625	7.958	-
(0.679)	137.520	152.069	28.941	146.000	418.605	190.909	138.706	191.143	329.760	-
	33.073	**41.347**	**15.268**	**67.449**	**73.913**	**53.452**	**63.316**	**84.899**	**77.514**	-
	32.120	38.472	9.003	38.061	76.543	43.748	39.170	47.211	70.632	-
	97.12	93.05	58.97	56.43	103.558	81.85	61.87	55.61	90.77	-
SnP	5.918	5.671	14.553	2.979	36.000	14.400	11.446	29.625	14.842	-
(1.000)	190.909	196.941	137.520	418.605	146.000	115.024	216.706	191.143	75.818	-
	41.598	**57.964**	**43.678**	**98.142**	**71.149**	**60.460**	**44.714**	**99.383**	**51.898**	-
	34.351	49.721	52.753	107.467	42.784	31.719	36.472	67.176	32.541	-
	82.58	85.78	120.78	109.50	60.13	52.463	81.57	67.59	62.70	-
SvlP	14.400	5.748	2.979	-	-	-	-	-	-	-
(0.536)	191.143	418.605	152.069	-	-	-	-	-	-	-
	69.145	**68.427**	**30.545**	-	-	-	-	-	-	-
	40.520	66.993	31.805	-	-	-	-	-	-	-
	58.60	97.90	104.12	-	-	-	-	-	-	-
SvnP	14.842	11.446	24.923	21.014	2.979	5.671	5.918	114.475	30.408	-
(0.533)	191.143	115.024	190.909	329.760	37.161	160.364	117.692	418.605	216.706	-
	74.598	**45.909**	**56.196**	**116.610**	**15.118**	**42.759**	**25.395**	**230.481**	**107.819**	-
	47.895	27.405	34.597	79.733	10.499	38.928	22.798	145.253	56.869	-
	64.21	59.69	61.57	68.38	69.45	91.04	89.77	63.02	52.74	-
SlnP	16.579	2.979	5.748	5.671	13.143	7.958	29.625	14.170	14.400	36.000
(0.528)	329.760	28.941	160.364	137.520	158.897	216.706	191.143	418.605	133.043	146.000
	96.280	**15.243**	**48.483**	**33.435**	**45.370**	**68.327**	**84.899**	**98.333**	**60.432**	**74.574**
	97.815	8.718	40.707	31.327	35.784	58.150	47.211	87.608	36.312	42.372
	101.59	57.19	83.96	93.70	78.87	85.11	55.61	89.09	60.09	56.82
SvlnP	13.143	2.979	-	-	-	-	-	-	-	-
(0.500)	191.143	418.605	-	-	-	-	-	-	-	-
	62.206	**55.957**	-	-	-	-	-	-	-	-
	43.000	62.792	-	-	-	-	-	-	-	-
	69.13	112.21	-	-	-	-	-	-	-	-

Note: from up to down, in each cell of the table are the minimum, maximum, mean, standard deviation values and the coefficient of variation of the target variable; mean values are given in bold, and C1 to C10 stand for the clustering solutions.

When using the performance metrics such as the efficiency (**Figure 3b**) or productivity (**Figure 3c**), the differentiation in mean values was higher, particularly when using the productivity as a target variable (**Figure 3c**). In **Figure 3**, the mean values of work cycle time (*T*), efficiency (*E*), and productivity (*P*) are plotted against the clusters ordered incrementally based on the mean values of these target variables. As a consequence, the dots aligned vertically for a given value in work cycle time (*T*), efficiency (*E*), or productivity (*P*) indicate small or no differences among the mean values of the corresponding clusters.

For instance, taking as an example the *SlnT* scenario, which was used to cluster the work cycle time as a function of piece length and number of recovered logs, the final solution consisted of ten clusters. However, a clear differentiation in the mean values of work cycle time was found only in the ranges of mean values from 0.021 to 0.026 hours (first three clusters) and from 0.026 to 0.033 hours (the last three clusters), leaving four clusters with similar mean values of work cycle time (0.026 hours). Similar observations are valid for the rest of scenarios characterized by a higher number of clusters (nine clusters), while a clearer differentiation of the mean values was found only for the scenarios that had a two- or a three-cluster solution.

Clustering scenarios based on the efficiency as a target variable were selected for the analysis of the way in which the feature variables were clustered. The choice was based on the results shown in **Figure 3b**, indicating a good differentiation between the mean values as coming from each cluster. The results characterizing the clustering of feature variables are shown in **Appendix A** for the seven scenarios taken into analysis.

Using the input volume (*v*) as a feature for clustering returned two clusters, that were formed based on well separated values of this feature variable. Mean values of efficiency were also well separated but the data ranges from which they were computed were overlapped. The rest of single feature variable-based clustering solutions (*SlE* - clustering by taking the piece length as a feature, *SnE* - clustering by taking the number of recovered logs as a feature) generally provided well separated groups of values in the feature variables such as the piece length (*l*) and the number of recovered logs (*n*).

As the number of feature variables used increased, some of them were not well separated in the clustering solution. For instance, using three feature variables (*l*, *v*, *n*) returned a clustering solution indicating two groups (**Appendix A**). In this solution, piece length (*l*) was well separated as opposed to the input volume (*v*) and number of recovered logs (*n*). **Appendix A** shows the results on feature variables as box plots for all the feature variables used in the seven clustering scenarios having as a target variable the efficiency.

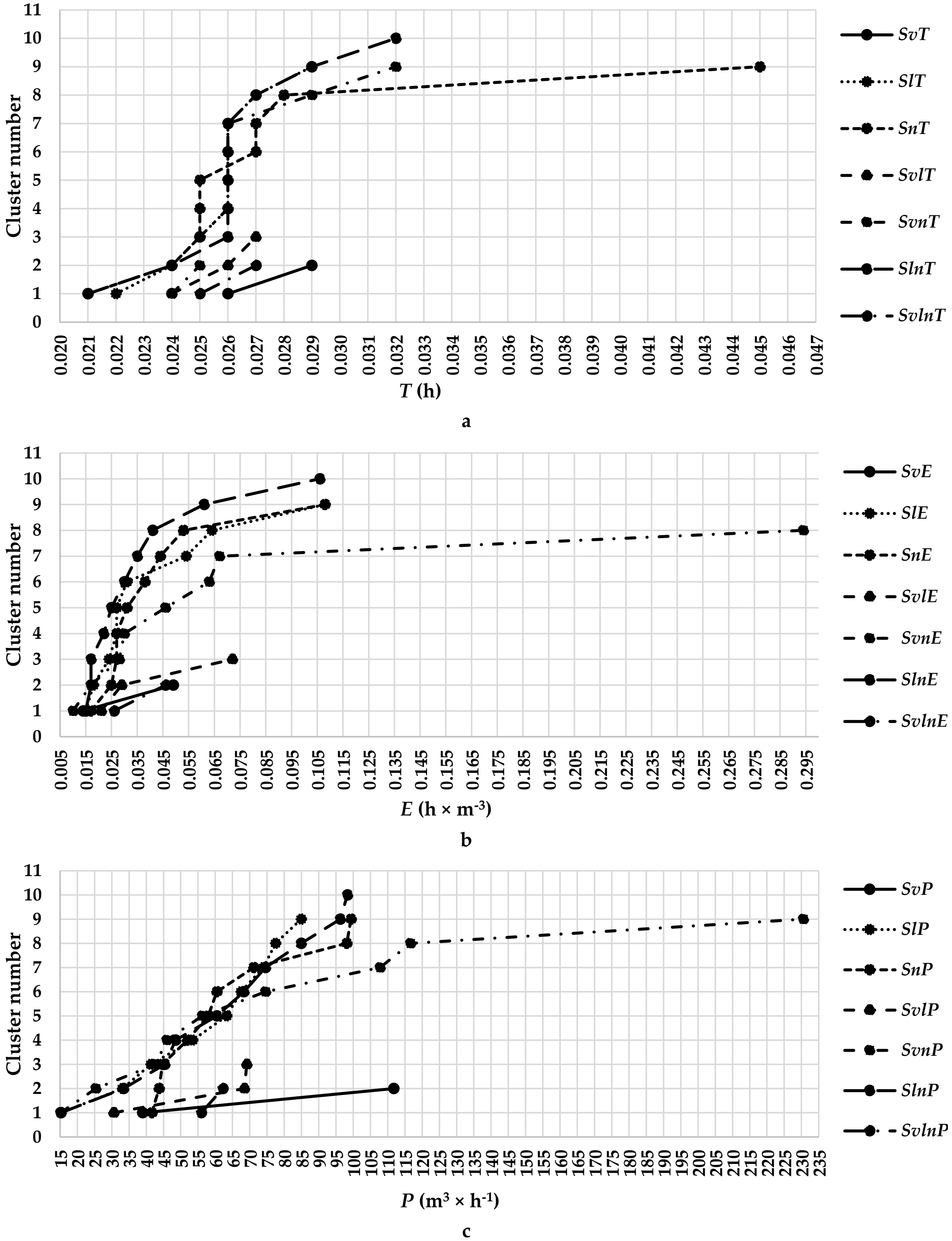


Figure 3. Average values of work cycle time (a), efficiency (b) and productivity (c) as clustered on scenarios by the k-means method. Note: the clusters were reordered based on incremental values of target variables.

4. DISCUSSION

The expectations of this study were to identify well differentiated data groups based on clustering the work cycle time, efficiency and productivity taking as features the input volume, piece length and number of recovered logs. Unfortunately, no similar studies were identified so as to provide a basis for comparison of the results. Depending on the number of features used, however, the solutions were more or less complex in terms of number of clusters, which ranged from two to ten. However, the number of clusters was the same when using the same set of feature variables, irrespective of the target variable in question. This may be due to the fact that work cycle time was used either as a target variable, or to compute the performance metrics (efficiency and productivity) which were then used as target variables.

The clustering solutions based on a single feature generally provided a good differentiation in the values of the features used, which avoided overlaps in their data range. However, they did not provide a good separation in the data range of target variables, excepting the case when the average values used to differentiate were computed from the clustered data. It seems that using the number of recovered logs as a feature variable was the best option in providing well differentiated clusters by the silhouette score, which probably comes from the fact that the feature variable used had discrete values. However, the target variables were not well differentiated in the resulted clusters.

Based on the results of the study, the mean values of the target variables characterizing the performance metrics (i.e., efficiency or productivity) can be used as descriptors of performance increment and for developing piece-rate systems assuming that a clear differentiation would be present in the feature variables used to cluster the data.

All of the two-cluster solutions have provided such a differentiation in the feature variable used, as well as in the mean values of the target variables, therefore they may qualify as a solution to segment and sufficiently differentiate in data. It is likely, however, that such a clustering solution will provide less differentiation in the target variables and will affect the dynamics in economic performance by missing a significant part of the data categorization potential. When using two features to cluster the data, it was common to find only one of the them as being well differentiated in groups by considering the data range. For instance, in the case of *SlvE* (clustering the efficiency as a target variable based on piece length and input volume as feature variables) scenario, the piece length provided a good differentiation in the data ranges which held true also for the scenario in which three features were used - *SvlnE* (clustering the efficiency as a target variable base on all feature variables, **Appendix A**). This indicates that such well differentiated features may be used as the main descriptors for the performance metrics used in a piece-rate system at the expense of omitting the rest of less informative feature variables.

k-means clustering method is just one of the many tools that can be used to group the data by an unsupervised approach, being useful in discovering patterns or groups of similar features in data. Future studies could evaluate the eventual improvements brought in data clustering based on the use of other statistical clustering techniques, as well as how using other settings to run the algorithm could improve the outcomes in terms of differentiation in data. This study, on the other hand, is based on a dataset of a limited size, a characteristic which may affect the outcomes in terms of clustering performance. How the size of the dataset may influence the quality of clustering is another topic that should be explored by future studies.

5. CONCLUSIONS

Based on the results of this study, the following may be concluded:

1) Using the k-means clustering technique to find well-differentiated groups in productivity data works well assuming that a single feature is used and that the mean values of the target variables (clustered data) are used as outcomes;

2) Increasing the number of feature variables used for clustering leads to a poorer differentiation in some of them although the mean values of the target variables are still properly differentiated when performance metrics such as the efficiency or productivity are used as targets;

3) Clustering solutions based on discrete feature variables provide better clustering solutions when considering the silhouette score as a metric to evaluate the goodness of a clustering solution, while the data range of the feature variables may affect the number of clusters in a given solution.

SUPPLEMENTARY MATERIALS

Not the case.

FUNDING

This work received no funding.

ACKNOWLEDGEMENTS

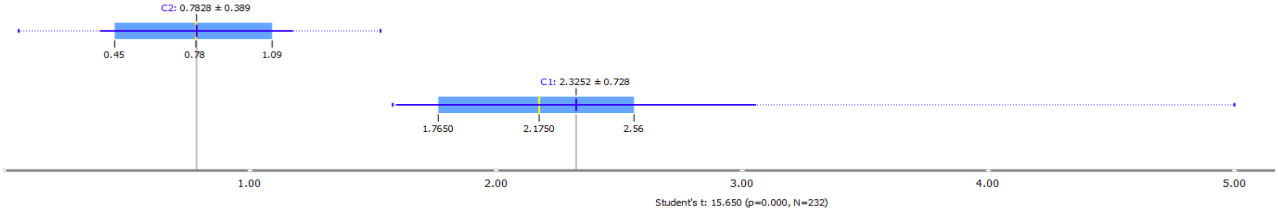
The Author would like to thank to the Department of Forest Engineering, Forest Management Planning and Terrestrial Measurements, Faculty of Silviculture and Forest Engineering, Transilvania University of Brasov, for providing the infrastructure needed for the study. This study uses the initial database that served for the modelling study of Borz S.A., Seceleanu V.N., Iacob L.M. & Kaakurivaara N., 2023. Bucking at landing by a single-grip harvester: fuel consumption, productivity, cost and recovery rate. Forests, 14, 465.

CONFLICT OF INTEREST

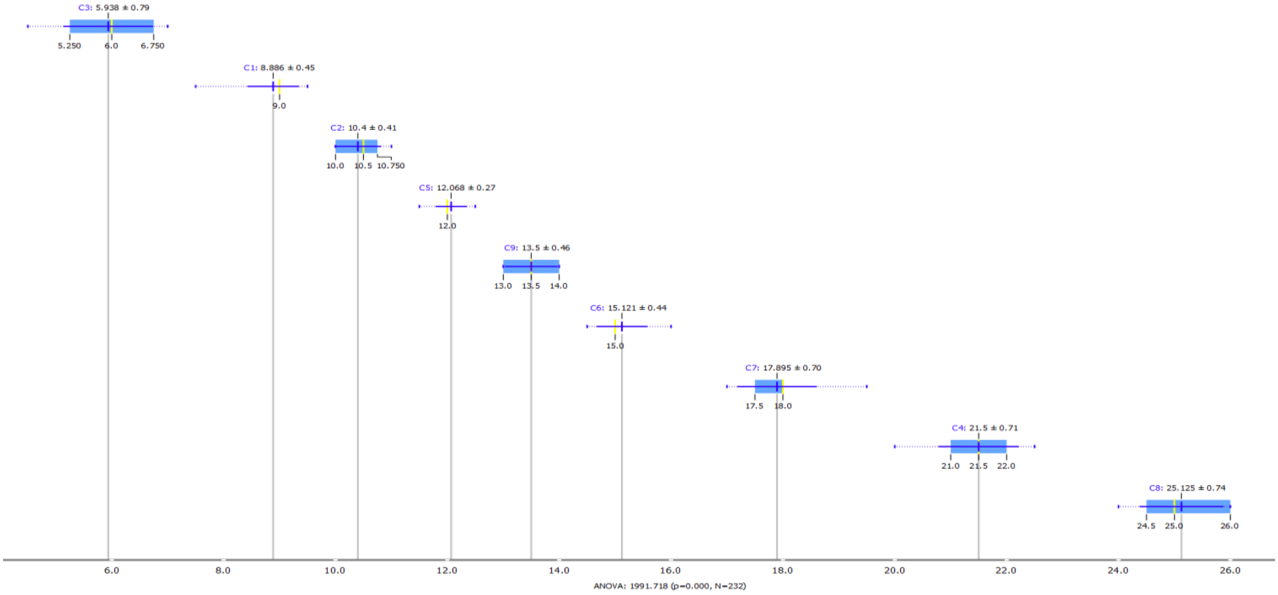
The Author declares no conflict of interest.

APPENDIX

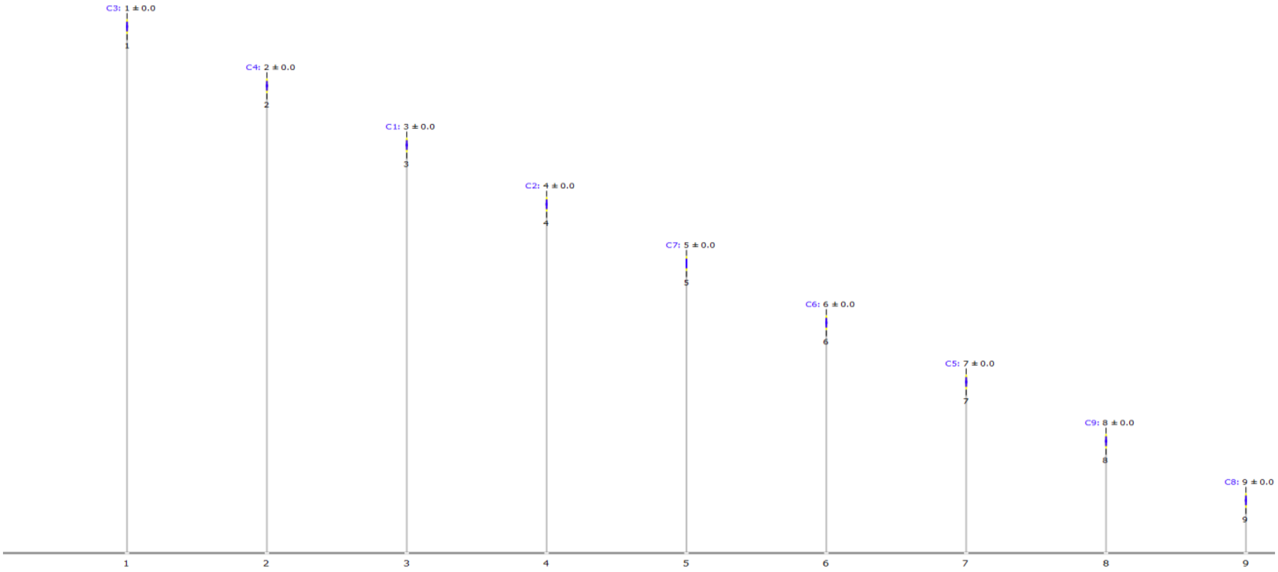
APPENDIX A – Descriptive statistics of feature variables on scenarios for efficiency set as target.



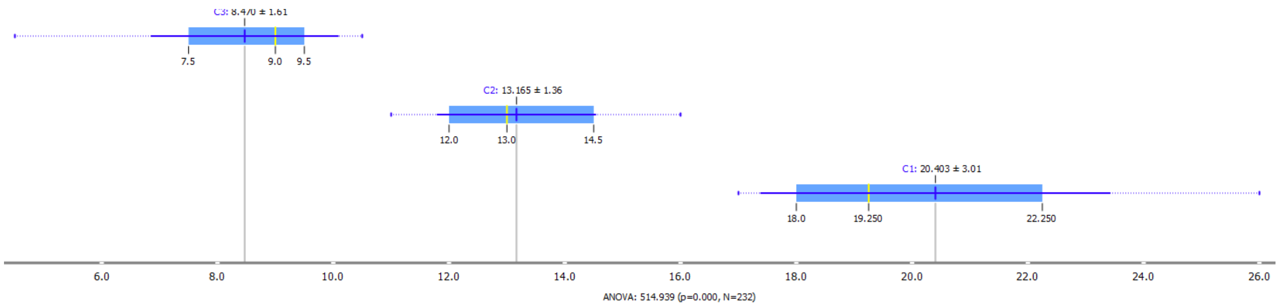
Scenario: *SvE*, feature variable - input volume (*v*)



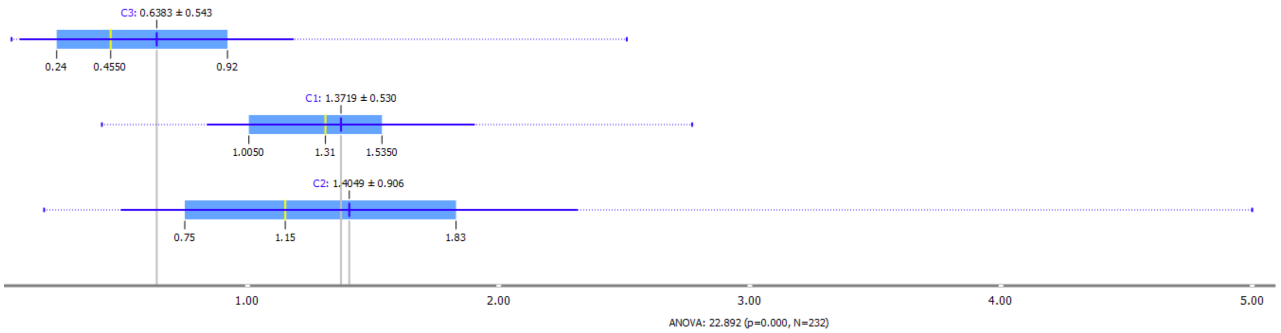
Scenario: *SlE*, feature variable - piece length (*l*)



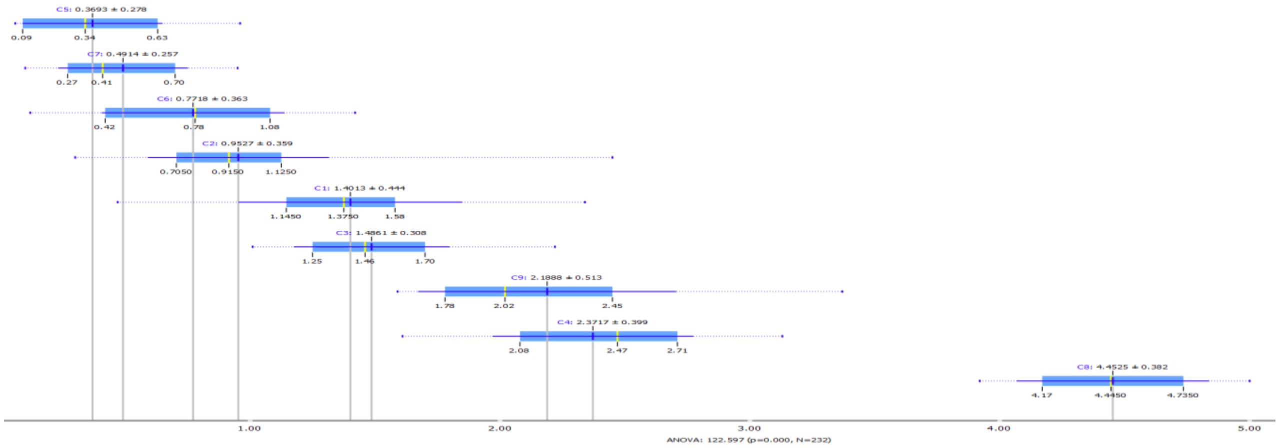
Scenario: *SnE*, feature variable - number of recovered logs (*n*)



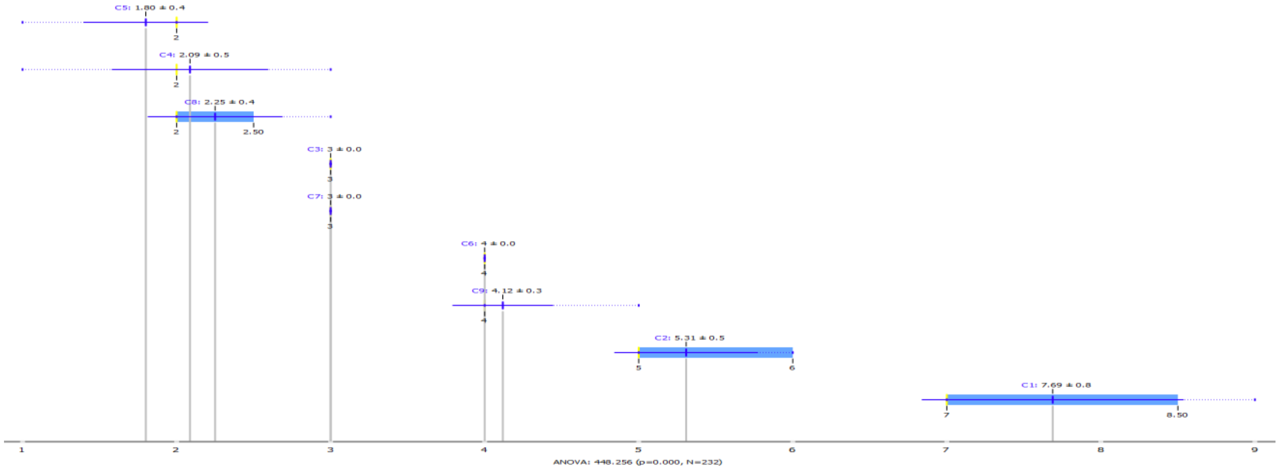
Scenario: *SlvE*, feature variable - piece length (*l*)



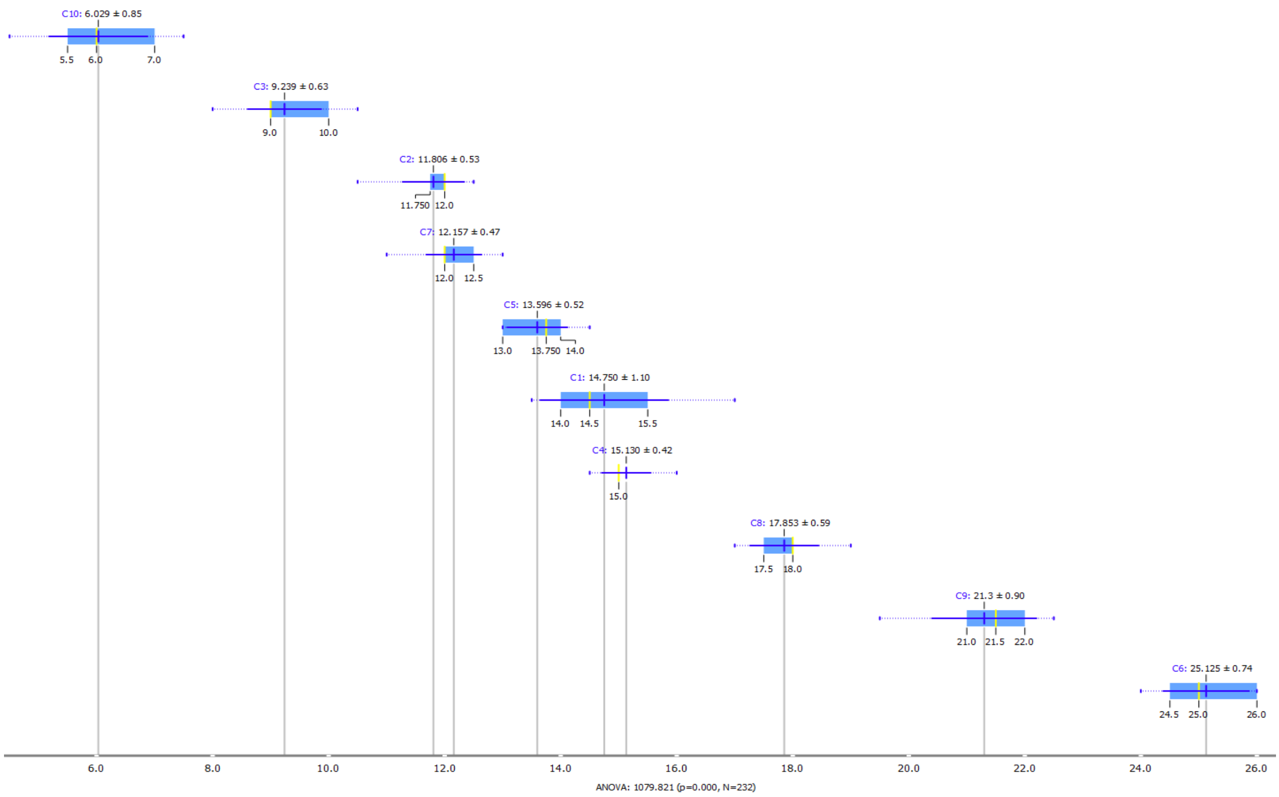
Scenario: *SlvE*, feature variable - input volume (*v*)



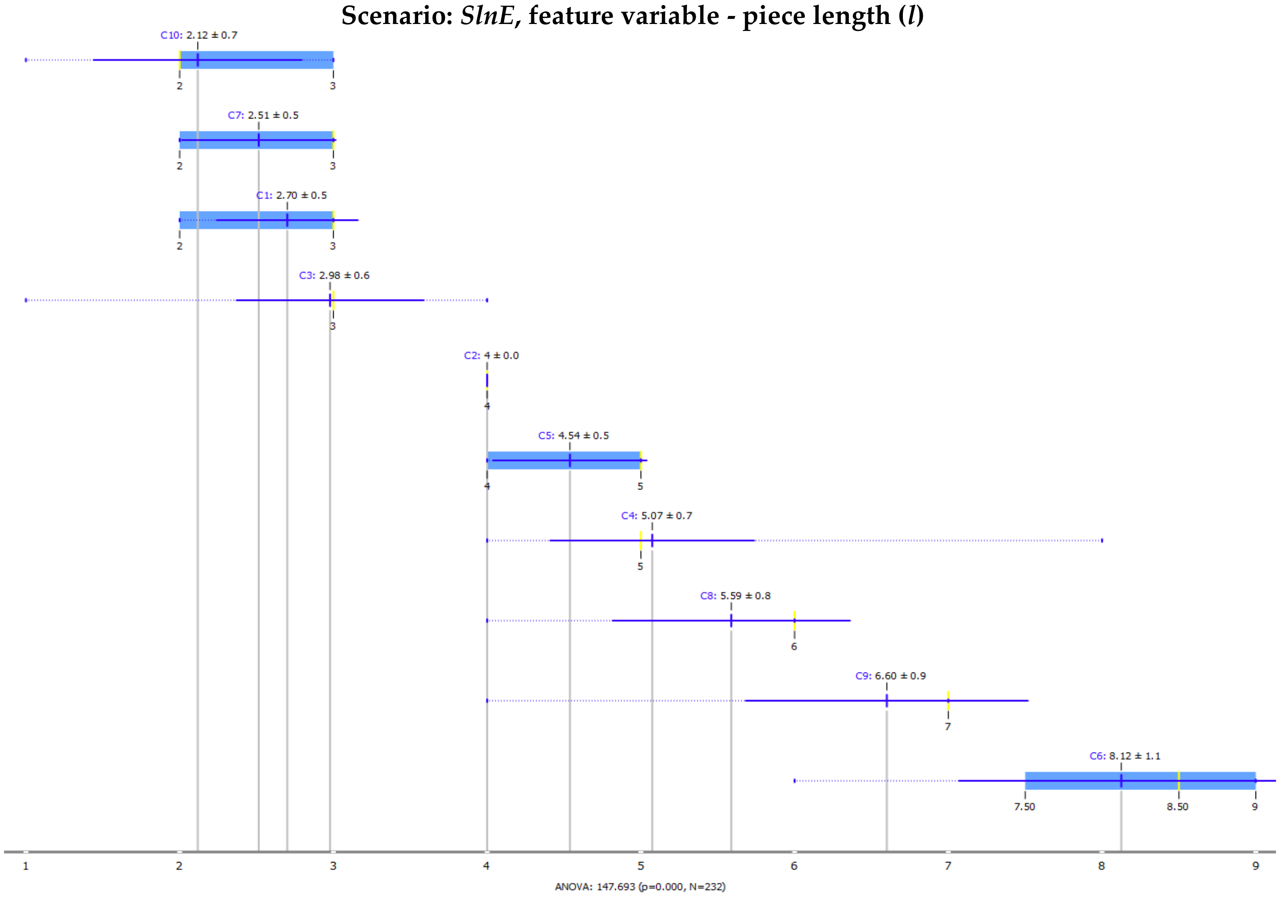
Scenario: *SvnE*, feature variable - input volume (*v*)



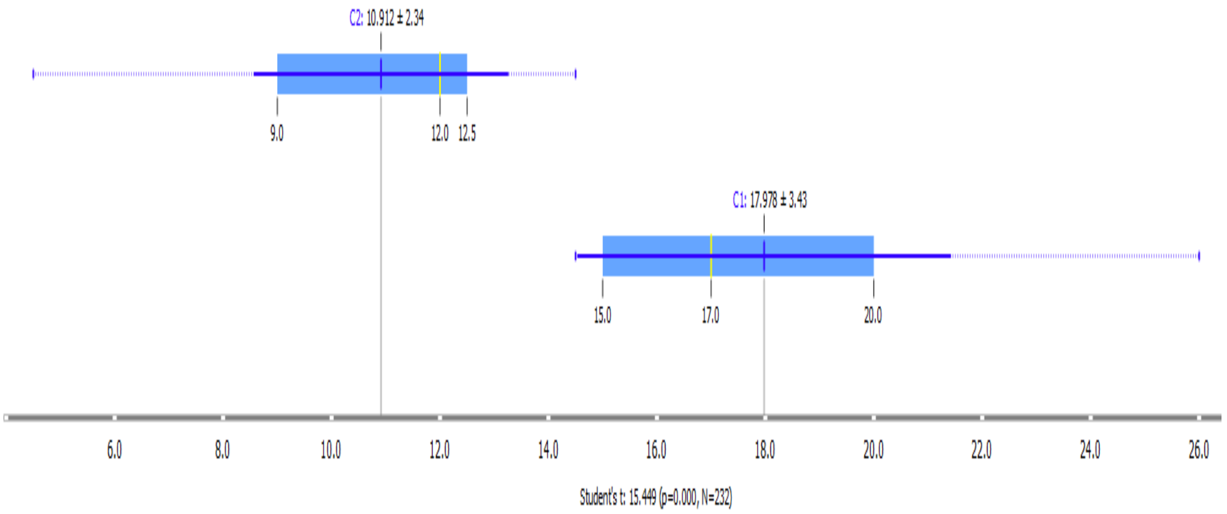
Scenario: *SvnE*, feature variable - number of recovered logs (*n*)



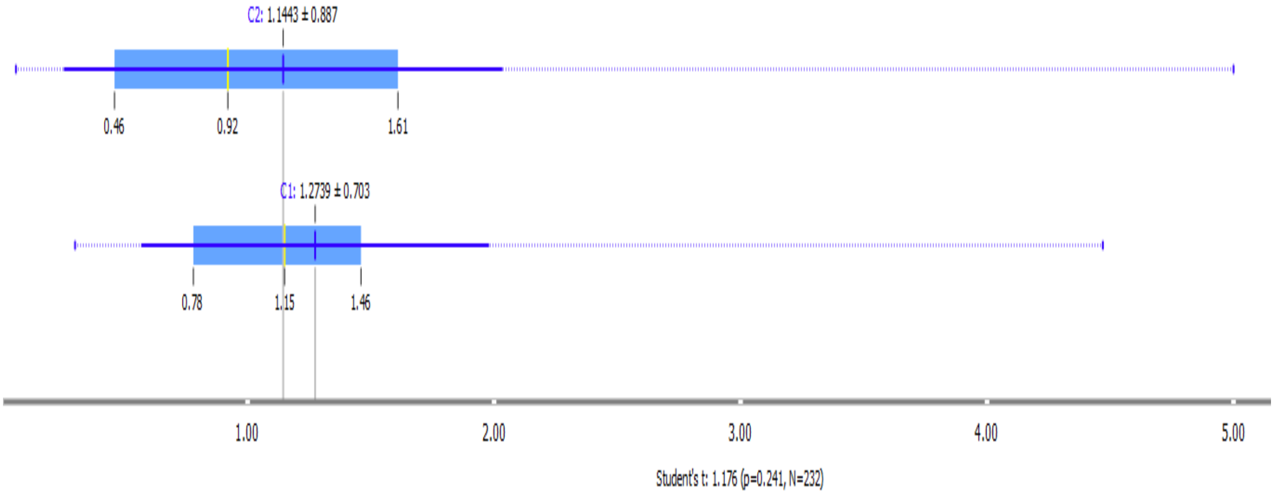
Scenario: *SlnE*, feature variable - piece length (*l*)



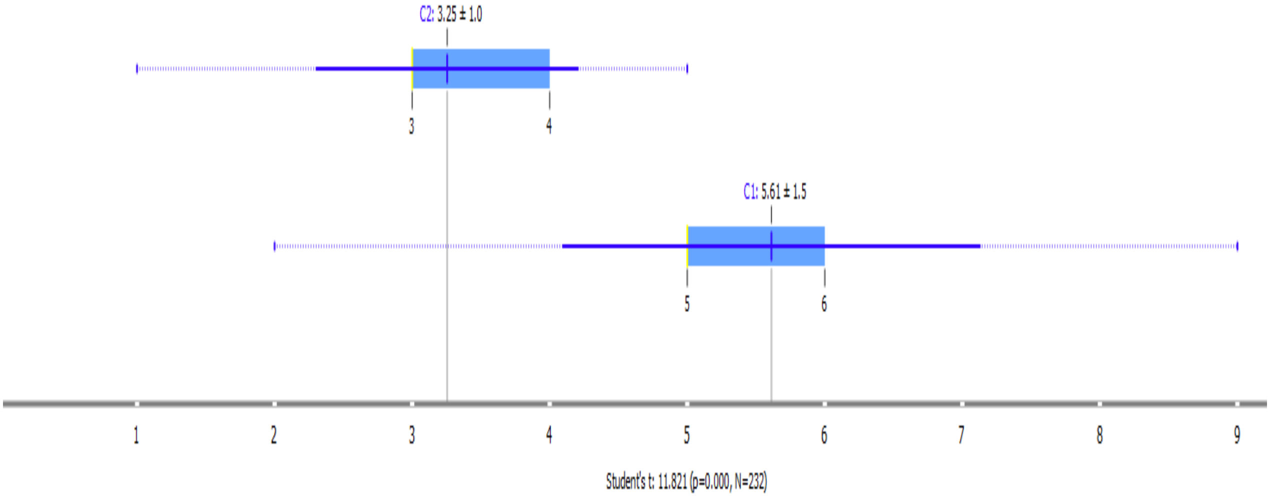
Scenario: *SlnE*, feature variable - number of recovered logs (*n*)



Scenario: *SvlnE*, feature variable - piece length (*l*)



Scenario: *SvlnE*, feature variable - input volume (*v*)



Scenario: *SvlnE*, feature variable - number of recovered logs (*n*)

EXTENDED ABSTRACT – REZUMAT EXTINS

Titlu în română: *Performanța algoritmului k-means în identificarea de grupuri de date omogene: studiu de caz cu privire la performanța operațională a unei mașini multifuncționale de recoltare folosită la fasonat pe platforma primară*

Introducere: *Creșterea eficienței economice în exploatarea lemnului a condus la creșterea gradului de mecanizare specific aceste activități, iar mașinile multifuncționale de recoltare sunt folosite în multe țări europene, inclusiv în România. Aceste mașini sunt folosite, în mod obişnuit, pentru doborârea, curățarea de crăci şi secţionarea arborilor. Din rațiuni legate nelimitativ de costurile operaționale şi securitatea muncii, în anumite situații s-a trecut la folosirea acestora şi pentru fasonarea lemnului pe platformele primare. Atunci când se operează cu catarge sau cu piese de lemn lung deramificate, ele servesc operației de secționare. Studiile realizate până în prezent au avut scopul de a modela consumul de timp şi productivitatea muncii în funcţie de variația unor factori operaționali, ca premisă pentru estimarea costurilor. Variabilitatea unor factori operaționali cum ar fi mărimea pieselor secționate şi decizia cu privire la locurile de secţionare nu oferă întotdeauna premisele obținerii unor modele predictive suficient de precise, motiv pentru care se poate recurge la gruparea datelor similar normelor de timp şi producție. Studiul de față testează măsura în care se poate folosi algoritmul k-means de grupare nesupervizată a datelor pentru a obține categorii bine diferențiate sub raportul unor variabile caracterizând performanța productivă şi factorii operaționali.*

Materiale şi metode: *Studiul are la bază un set de date conținând consumul de timp la nivel de fază pentru un număr de peste 230 de cicluri de muncă caracterizând secţionarea cu o mașină multifuncțională de recoltare. Pentru fiecare observație din setul de date au fost disponibile valorile cu privire la volumul pieselor intrate în operație, lungimea acestora şi numărul de piese rezultate după secționare, precum şi date cu privire la duratele fazelor din ciclurile de muncă. Având la bază aceste date s-au calculat productivitatea ($m^3 \times h^{-1}$) şi eficiența ($h \times m^{-3}$) pentru fiecare ciclu de muncă, apoi s-au creat scenarii de grupare a datelor (21 de scenarii) care au luat în considerare variabilele caracterizând factorii operaționali (volumul piesei, lungimea piesei şi numărul de piese rezultate prin secționare) şi indicatorii de performanță (consumul de timp al unui ciclu de muncă, eficiența şi productivitatea muncii), astfel încât să se acopere toate combinațiile posibile de factori operaționali. Aceste scenarii au fost folosite pentru a grupa datele prin folosirea algoritmului k-means, pas care s-a realizat în programul Orange Visual Programming prin setarea numărului posibil de grupuri de date între 2 şi 10 şi a numărului de iterații la 10000. Calitatea generală a grupării s-a evaluat prin folosirea unui indicator specific („silhouette score") iar calitatea grupării datelor în categorii s-a evaluat prin modul în care s-au suprapus amplitudinile de variație caracterizând grupurile rezultate.*

Rezultate şi discuții: *Soluțiile obținute au conținut între două şi zece categorii, iar numărul de categorii a fost același pentru un anumit set de variabile operaționale utilizate în analiză. Nicio soluție nu a furnizat o separare clară a valorilor indicatorilor de performanță prin luarea în considerare a domeniului de variație al valorilor specifice. Prin folosirea valorilor medii obținute din datele grupate în categorii, rezultatele s-au îmbunătățit mai ales în cazul eficienței şi productivității. În măsura în care numărul de variabile caracterizând condiţiile operaționale a crescut, valorile unora dintre acestea nu au mai fost bine separate în categoriile rezultate. Soluțiile care au avut la bază o singură variabilă operațională au furnizat o separare bună a valorilor acesteia dar nu şi a valorilor indicatorilor de performanță. Având la bază rezultatele acestui studiu, valorile medii ale categoriilor cu privire la eficiență şi productivitate pot fi utilizate ca descriptori ai creșterii performanței presupunând că există o diferențiere clară a valorilor variabilelor operaționale. Studii ulterioare pot să clarifice modul în care alte metode de grupare a datelor pot să îmbunătățească calitatea separării sau modul în care folosirea altor setări pentru același algoritm poate să conducă la o diferențiere mai bună a datelor.*

Concluzii: *Utilizarea algoritmului k-means pentru a identifica categorii bine diferențiate cu privire la performanța operațională produce rezultate bune atunci când se folosește o singură variabilă operațională concomitent cu folosirea ca descriptori ai performanței a valorilor medii din categoriile rezultate. Creșterea numărului de variabile operaționale conduce la o diferențiere mai slabă a unora dintre acestea.*

Cuvinte cheie: *exploatarea lemnului, eficiență, diferențiere, k-means, normare.*

REFERENCES

1. Moskalik T., Borz S.A., Dvorák J., Ferencik M., Glushkov S., Muiste P., Lazdinš A., Styranivsky O., 2017. Timber harvesting methods in Eastern European countries: a review. Croatian Journal of Forest Engineering, 38, 231-241.

2. Zurita Vintimilla M.C., Castro Perez S.N., Borz S.A., 2021. Processing small-sized trees at landing by a double-grip machine: a case study on productivity, cardiovascular workload and exposure to noise. Forests, 12, 213.

3. Borz S.A., Seceleanu V.N., Iacob L.M., Kaakurivaara N., 2023. Bucking at landing by a single-grip harvester: fuel consumption, productivity, cost and recovery rate. Forests, 14, 465.

4. Oprea I., 2008. Tehnologia exploatării lemnului. Editura Universității Transilvania din Braşov, 273 p.

5. Apăfăian A.I., Proto A.R., Borz S.A., 2017. Performance of a mid-sized harvester-forwarder system in integrated harvesting of sawmill, pulpwood and firewood. Annals of Forest Research, 60, 227-241.

6. Visser R., Spinelli R., 2011. Determining the shape of the productivity function for mechanized felling and felling-processing. Journal of Forest Research, 17(5), 397-402.

7. Björheden R., Apel K., Shiba M., Thompson M., 1995. IUFRO Forest Work Study Nomenclature. Department of Operational Efficiency, Swedish University of Agricultural Science, Grapenberg, Sweden, 16 p.

8. Borz S.A., 2008. Evaluarea eficienţei echipamentelor şi sistemelor tehnice în operaţii forestiere. Editura Lux Libris, Braşov, România, 252 p.

9. Acuna M., Bigot M., Guerra S., Hartsough B., Kanzian C., Kärhä K., Lindroos O., Magagnotti N., Roux S., Spinelli R., et al., 2012. Good practice guidelines for biomass production studies. Magagnotti N., Spinelli R., Eds.; CNR IVALSA: Sesto Fiorentino, Italy.

10. Zinkevicius R., Steponavicius D., Vitunskas D., Cinga G., 2012. Comparison of harvester and motor-manual logging in intermediate cuttings of deciduous stands. Turkish Journal of Agriculture and Forestry, 36, 591-600.

11. Borz S.A., Bîrda M., Ignea G., Popa B., Câmpu V.R., Iordache E., Derczeni R.A., 2014. Efficiency of a Woody 60 processor attached to a Mounty 4100 tower yarder when processing coniferous timber from thinning operations. Annals of Forest Research, 57, 333-345.

12. Nurminen T., Korpunen H., Uusitalo J., 2006. Time consumption analysis of the mechanized cut-to-length harvesting system. Silva Fennica, 40, 335-363.

13. Väätäinen K., Ala-Fossi A., Nuutinen Y., Roser D., 2006. The effect of single grip harvester's log bunching on forwarder efficiency. Baltic Forestry, 12, 64-69.

14. Eriksson M., Lindroos O., 2014. Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. International Journal of Forest Engineering, 25, 179-200.

15. Mederski P.S., Bembenek M., Karaszewski Z., Lacka A., Szczepanska-Alvarez A., Rosinska M., 2016. Estimating and modelling harvester productivity in pine stands of different ages, densities and thinning intensities. Croatian Journal of Forest Engineering, 37, 27-36.

16. Norihiro J., Ackerman P., Spong B.D., Langin D., 2018. Productivity model for cut-to-length harvester operation in South African Eucalyptus pulpwood plantations. Croatian Journal of Forest Engineering, 39, 1-13.

17. Glöde D., 1999. Single- and double-grip harvesters: Productive measurements in final cutting of shelterwood. International Journal of Forest Engineering, 10, 63-74.

18. Heinimann H.R., 2007. Forest operations engineering and management - the ways behind and ahead of a scientific discipline. Croatian Journal of Forest Engineering, 28, 107-121.

19. Marchi E., Chung W., Visser R., Abbas D., Nordfjell T., Mederski P.S., McEwan A., Brink M., Laschi A., 2018. Sustainable forest operations (SFO): a new paradigm in a changing world and climate. Science of the Total Environment, 634, 1385-1397.

20. Ministerul Industrializării Lemnului şi Materialelor de Construcții. Centrala de Exploatare a Lemnului București, 1989. Norme şi normative de muncă unificate în exploatările forestiere, 495 p.

21. Lloyd S.P., 1957. Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.

22. Demsar J., Curk T., Erjavec A., Gorup C., Hocevar T., Milutinovic M., Mozina M., Polajnar M., Toplak M., Staric A., et al., 2013. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, 14, 2349-2353.

23. Rousseeuw P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.