Growth Curve Models and Clustering Techniques for the Study of New Sugarcane Hybrids: An Integrated Approach with K-Means, K-Medoids, and DBSCAN

Carlos David Carretillo Moctezuma , María Guzmán Martínez * , Flaviano Godínez Jaimes , José Concepción García Preciado , Ramón Reyes Carreto , José Torrones Salgado , Edgar Pérez Arriaga

Posted Date: 3 September 2024

doi: 10.20944/preprints202409.0179.v1

Keywords: Growth curves; clustering; varieties; sucrose percentage

 doi:10.20944/preprints202409.0179.v1

Article

Growth Curve Models and Clustering Techniques for the Study of New Sugarcane Hybrids: An Integrated Approach with K-Means, K-Medoids, and DBSCAN

Carlos David Carretillo Moctezuma [1], María Guzmán Martínez [1,*], Flaviano Godínez Jaimes [1], José Concepción García Preciado [2], Ramón Reyes Carreto [1], José Terrones Salgado [3] and Edgar Pérez Arriaga [4]

[1] Facultad de Matemáticas, Universidad Autónoma de Guerrero, Av. Lázaro Cárdenas SN CP 39087, Chilpancingo de los Bravo, Guerrero
[2] Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Campo Experimental Tecomán, Tecomán, Colima, México
[3] Decanato de Ciencias de la Vida y la Salud, Escuela de Ingeniería en Agronomía, Centro de Investigación en Horticultura y Plantas Nativas, UPAEP University
[4] Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Victoria, División de Estudios de Posgrado e Investigación
* Correspondence: marnezmar@yahoo.com.mx; Tel.: +747 256 9489

Abstract: Sugarcane (*Saccharum spp.*) is a crop of great industrial and alimentary importance, essential for the production of numerous products. Given its significance, genetic improvement programs exist that involve a rigorous study process, from material selection to the development of new varieties, requiring at least seven selection phases. This study modeled the growth curve of the sucrose percentage (SP) in 33 hybrids and six control varieties (MEX 69-290, ITV 92-1424, CP 72-2086, COLMEX 94-8, COLMEX 95-27, RB 85-5113) during the plant and ratoon periods in the experimental fields of the Melchor Ocampo Sugar Mill, Jalisco, Mexico. For clustering the materials, k-means, k-medoids, and Density-based spatial clustering of applications with noise (DBSCAN) algorithms were used, considering four maturity types among the control varieties. The DBSCAN algorithm proved to be the most effective, as the means between groups were not statistically equal. The hybrids identified as candidates for subsequent phases due to their high SP were COSTA JAL, ATEMEX 99-48, ATEMEX 99-1, ATEMEX 99-61, MEX 70-486, MEX 80-1521, ITSAMEX 07-44814, ITSAMEX 06-6395, and ITSAMEX 07-1903. These results are crucial for improving the productivity and sustainability of the crop, with significant implications for the sugar industry.

Keywords: growth curves; clustering; varieties; sucrose percentage

1. Introduction

The development of new sugarcane varieties offers a solution to the genetic deterioration experienced by commercial varieties that have been cultivated for several years and tend to exhibit a loss in their genetic diversity. This deterioration becomes evident through a decrease in the field yield, which subsequently impacts the industrial use of sugarcane. The lack of new genotypes to replace the deteriorated commercial varieties is the primary factor affecting the yield [1,2].

To obtain new varieties of sugarcane, materials that respond optimally to the environmental conditions and the needs of sugar agro-industries are proposed. The materials must meet requirements such as resistance to pests, diseases, and prolonged drought periods, as well as good adaptation to soil types, low post-harvest deterioration, and reduced production costs. It is also necessary for these varieties to have a good juice quality and an optimal amount of sucrose concentration. These traits will define the new commercial varieties [3,4].

Developing an outstanding commercial sugarcane variety for a specific region takes approximately ten to twelve years of constant evaluation both in the field and in an industrial setting. This process begins with the production of the seedling and ends with the evaluation of its performance in the field. The main phases that new materials must go through to become a commercial variety include strain selection, furrow testing, multiplication I, multiplication II, multiplication III, plot establishment,

 doi:10.20944/preprints202409.0179.v1

adaptation, agro-industrial evaluation, and semi-commercial testing [3,5].

During the adaptability phase, the goal is to evaluate the new hybrids under different soil and climate conditions. In this stage, both agronomic and industrial quality variables are assessed. The main agronomic variables include the germination percentage, height, population density, and health status. On the other hand, industrial quality variables include the sucrose percentage (SP), degrees Brix, purity, fiber content, and reducing sugars. Destructive sampling is carried out during this phase to study the maturity curves of the genotypes [5,6].

In the agro-industrial phase, the most important variables include the polarization percentage (PP), which indicates the content of sucrose in grams in the sample [7], degrees Brix, purity, fiber content, juiciness, and hardness of the sugarcane rind [5]. Some researchers utilize the PP to study the maturity of the materials [8] since the apparent sucrose is equivalent to the PP [9].

The physiological or biological maturity of sugarcane is determined by the accumulation of sucrose over time [10]. This leads to what is known as the sucrose accumulation curve, which can help one to identify the maturity type of the sugarcane cultivar. This curve is of interest to the sugar industry for genotype selection and is studied starting in the material's adaptability phase [5,11].

Growth curves in agricultural crops are useful for understanding a plant's development over time within a specific ecosystem. Typically, studying these curves involves making repeated measurements over time, and such data are referred to as longitudinal data [12]. Since this type of data often exhibits temporal autocorrelation, it is essential to employ an appropriate statistical methodology to analyze these data.

The growth curve of the SP is modeled using linear models with mixed effects that capture the variability of genotypes over time and the variability between genotypes. These mixed-effects models are known as growth curve models and are suitable for studying long yield cycles in perennial plants [13].

Some researchers compared commercial and new hybrids using measurements over time of several variables such as the degrees Brix, the PP, and the accumulation of sucrose. Quadratic regression was used to determine the sucrose accumulation over time, identify the optimal harvesting time, and maximize the sugar yield from each cultivar [14].

In another study, sucrose accumulation curves for different sugarcane genotypes in the Tucumán region in Argentina were studied. The variable under study was the PP. For each genotype, a nonlinear regression model composed of two polynomials with a break point, also known as a segmented polynomial, was fitted. To cluster the materials, the authors used the estimated sucrose accumulation curve and applied hierarchical clustering and partitioning methods, specifically Unweighted Pair Group Method using Arithmetic averages (UPGMA) and k-means, respectively. They found that the sucrose accumulation curves of the genotypes are influenced by the environment [11].

Some studies have been conducted on sugarcane genotypes to determine the agronomic response of new cultivars. The variables assessed were the tons of sugarcane per hectare, the tons of PP per hectare, and the sugar content throughout the harvest period. Using analysis of variance, multivariate analysis, and linear regression, researchers determined the best agro-industrial response, the optimal maturity stage for these cultivars, and the classification of the maturity type into the categories early, middle, and late [8].

The materials studied in this research have not been examined beyond the adaptability phase; therefore, there is uncertainty regarding how they will perform in terms of SP compared to the established varieties. This study will enable us to comprehend the sucrose accumulation curve's growth pattern for both the new materials and the control varieties, as well as variations within each group across different cycles. This information could provide valuable insights for genetic improvement and the selection of more productive sugarcane varieties.

The objective of this work is to study the sucrose accumulation curves of 33 hybrids (new materials) that are in the adaptability testing phase, together with six control varieties, i.e., commercial varieties MEX 69-290, ITV 92-1424, CP 72-2086, COLMEX 94-8, COLMEX 95-27 and RB 85-5113, and to form groups that are classified according to sucrose content in the cane, divided into early, early-intermediate,

 doi:10.20944/preprints202409.0179.v1

intermediate-late and late-ripening types [15]. It is important to note that these materials have not been studied beyond the adaptability phase; therefore, there is uncertainty about how they will behave in terms of the SP compared to the established varieties. However, the employed methodology will provide a more precise description of their behavior. This study will allow us to understand the growth curve of the SP for the new and control materials, as well as the variations in these materials from one cycle to another; this could provide relevant information for genetic improvement and the selection of more productive varieties.

2. Materials and Methods

A total of 39 materials (Table 1) from the experimental fields of the National Institute of Forestry, Agricultural and Livestock Research (INIFAP) were studied. They form two groups: the control materials and the new materials. The control group corresponds to six commercial varieties and the new materials correspond to 33 varieties of Mexican and foreign origin. The new materials are in the adaptability phase, and all the materials were evaluated in the plant and ratoon crop cycles under irrigation conditions in both seasons.

The experiment took place at Las Pilas experimental field (19;46;48.85 N, 104;13;57.65 W) and the Melchor Ocampo Sugar Mill (19;47;18.0 N, 104;14;24.3 W), which is located in the Zacapala area of Autlán de Navarro, Jalisco, México [6]. The experimental field has loamy-sandy soil, with an accumulated annual precipitation of 732 mm, an altitude of 860 m above sea level, and an average annual temperature of 23.9 °C. The experimental unit consisted of four rows, with each row measuring 15.0 m in length and 1.4 m in width. The useful plot area consisted of the two central rows, with 2 m removed from the ends.

The SP of the materials was measured through destructive sampling on four dates (Table 2), with an approximate difference of one month between the measurement dates.Seven months elapsed from planting to the first measurement date in the plant cycle, and six months from harvesting to the first measurement date in the ratoon cycle.

Table 1. Materials evaluated during the plant and ratoon crop cycles.

ID	Material	Origin	Type	ID	Material	Origin	Type
1	COSTA JAL	Mexican	N	21	ITSAMEX 07-7259	Mexican	N
2	RB 85-5035	Foreign	N	22	ITSAMEX 06-6395	Mexican	N
3	ITV 92-1424*	Foreign	C	23	ITSAMEX 06-4863	Mexican	N
4	RB 85-5113*	Foreign	C	24	ITSAMEX 07-4954	Mexican	N
5	LAICA 92-13	Foreign	N	25	ITSAMEX 07-4387	Mexican	N
6	CP 72-2086*	Foreign	C	26	ITSAMEX 07-121115	Mexican	N
7	COLMEX 94-8*	Mexican	C	27	ITSAMEX 07-246	Mexican	N
8	ATEMEX 99-48	Mexican	N	28	ITSAMEX 07-12116	Mexican	N
9	ATEMEX 99-1	Mexican	N	29	ITSAMEX 07-12113	Mexican	N
10	ATEMEX 99-61	Mexican	N	30	ITSAMEX 07-2963	Mexican	N
11	MEX 70-486	Mexican	N	31	ITSAMEX 07-99711	Mexican	N
12	TCP 89-3505	Foreign	N	32	ITSAMEX 07-9886	Mexican	N
13	MEX 80-1521	Mexican	N	33	ITSAMEX 07-86810	Mexican	N
14	MEX 69-290*	Mexican	C	34	ITSAMEX 07-12119	Mexican	N
15	ITSAMEX 07-44814	Mexican	N	35	ITSAMEX 07-1903	Mexican	N
16	ITSAMEX 06-3049	Mexican	N	36	ITSAMEX 07-44813	Mexican	N
17	ITSAMEX 07-8681	Mexican	N	37	ITSAMEX 07-12118	Mexican	N
18	ITSAMEX 07-20810	Mexican	N	38	CP 85-1382	Foreign	N
19	ITSAMEX 07-7501	Mexican	N	39	COLMEX 95-27*	Mexican	C
20	ITSAMEX 07-1107	Mexican	N				

[1] ∗ Commercial variety. N= New material, C= Commercial

 doi:10.20944/preprints202409.0179.v1

Table 2. SP evaluation dates of the 39 materials studied.

Cycle	Period	Date 1	Date 2	Date 3	Date 4
Plant	Planting 03-04-2011	10-08-2011	11-08-2011	12-08-2011	01-11-2012
Ratoon	Harvest 05-21-2012	11-16-2012	12-19-2012	01-22-2013	02-25-2013

2.1. Growth curve model

The linear mixed model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ response vector; $\mathbf{X}$ and $\mathbf{Z}$ are design matrices of dimensions $n \times p$ and $n \times q$, respectively; $\boldsymbol{\beta}$ is the fixed-effect parameter vector of dimensions $p \times 1$; $\mathbf{u}$ is the random-effects vector of dimensions $q \times 1$; and $\boldsymbol{\epsilon}$ is the error vector of dimensions $n \times 1$. In Model (1), $\mathbf{X}\boldsymbol{\beta}$ is the fixed component, and $\mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ is the random component. It is assumed that $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are independent, that is, $Cov(\mathbf{u}, \boldsymbol{\epsilon}) = \mathbf{0}$, and that they each have a multivariate normal distribution:

$$\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{D}),$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{R}),$$

where $\mathbf{D}$ and $\mathbf{R}$ are the covariance matrices of the random-effects vector and the error vector, respectively. From these assumptions, we have

$$\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),$$

where $\mathbf{V} = Cov(\mathbf{y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}$. By using the Henderson equations [16] to estimate the Model (1) parameters, the best unbiased linear estimator for $\boldsymbol{\beta}$ and the best unbiased linear predictor for $\mathbf{u}$ are obtained:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$
$$\widehat{\mathbf{u}} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Let y_{ij} be the j-th observation of the i-th material, with $j = 1, ..., n_i$ (number of measurements) and $i = 1, ..., n$ (number of materials). The growth curve model is given by

$$y_{ij}(t_{ij}) = \beta_0 + \beta_1 t_{ij} + u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}. \tag{2}$$

The fixed component of Model (2) is $\beta_0 + \beta_1 t_{ij}$ and the random component is $u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}$. According to this model and considering four measurement dates, the design matrices for each material i are of dimensions 4×2 and are given by

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ 1 & t_{i3} \\ 1 & t_{i4} \end{pmatrix}.$$

 doi:10.20944/preprints202409.0179.v1

The fixed-effects parameter vector, the random-effects vectors, and the error vectors are given by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \ \mathbf{u}_i = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix}, \ \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \end{pmatrix}.$$

Then, the model for the i-th material is given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i, \ i = 1, ..., 39. \tag{3}$$

The random effects $\mathbf{u}_i$, $i = 1, ..., 39$, are assumed to be independent and identically distributed, that is, $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, with

$$\mathbf{D} = Var(\mathbf{u}_i) = \begin{pmatrix} \sigma^2_{u_{0i}} & \sigma_{u_{1i}u_{0i}} \\ \sigma_{u_{0i}u_{1i}} & \sigma^2_{u_{1i}} \end{pmatrix}.$$

There are several covariance structures for $\mathbf{D}$ that provide a more accurate representation of the variability and correlation between observations within a group or subject [17] that enables more flexible and appropriate model fitting to longitudinal data.

The error vectors $\boldsymbol{\epsilon}_i$, $i = 1, ..., 39$, are also assumed to be independent and identically distributed, that is, $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$, with

$$\mathbf{R}_i = Cov(\boldsymbol{\epsilon}_i) = \begin{pmatrix} Var(\epsilon_{i1}) & Cov(\epsilon_{i1}, \epsilon_{i2}) & Cov(\epsilon_{i1}, \epsilon_{i3}) & Cov(\epsilon_{i1}, \epsilon_{i4}) \\ Cov(\epsilon_{i1}, \epsilon_{i2}) & Var(\epsilon_{i2}) & Cov(\epsilon_{i2}, \epsilon_{i3}) & Cov(\epsilon_{i2}, \epsilon_{i4}) \\ Cov(\epsilon_{i1}, \epsilon_{i3}) & Cov(\epsilon_{i3}, \epsilon_{i2}) & Var(\epsilon_{i3}) & Cov(\epsilon_{i3}, \epsilon_{i4}) \\ Cov(\epsilon_{i1}, \epsilon_{i4}) & Cov(\epsilon_{i4}, \epsilon_{i2}) & Cov(\epsilon_{i4}, \epsilon_{i3}) & Var(\epsilon_{i4}) \end{pmatrix}.$$

This matrix can be decomposed in the form

$$\mathbf{R}_i = \sigma^2\boldsymbol{\Lambda}_i\mathbf{C}_i\boldsymbol{\Lambda}_i,$$

where $\boldsymbol{\Lambda}_i = diag(\lambda_1, ..., \lambda_{n_i})$ is a diagonal matrix with nonnegative diagonal elements for the i-th material, with λ_j, $j = 1, ..., n_i$, representing functions of variance [17]; $\mathbf{C}_i$ is a correlation matrix for the i-th material. This way, the presence of heteroscedasticity in the elements of the vector $\boldsymbol{\epsilon}_i$ is explained by $\boldsymbol{\Lambda}_i$, and $\mathbf{C}_i$ explains the correlation between the observations within the group. Some structures for $\mathbf{C}_i$ include autoregressive processes of order 1 and autoregressive structures [18].

For the plant cycle, Model (3) was used. The varPower variance function was used, indicating that the magnitude of the variance error changes nonlinearly over time. The specification of $\boldsymbol{R}_i$ assumes that $\mathbf{C}_i = \mathbf{I}$, and the elements of $\boldsymbol{\Lambda_i}$ are defined by the varPower $(\cdot)$ variance function; their dependence on time is given by $\lambda_j = |t_{ij}|^\delta$, where δ is a constant. Then,

$$\boldsymbol{R}_i = \sigma^2 \begin{pmatrix} t_{i1}^{2\delta} & 0 & 0 & 0 \\ 0 & t_{i2}^{2\delta} & 0 & 0 \\ 0 & 0 & t_{i3}^{2\delta} & 0 \\ 0 & 0 & 0 & t_{i4}^{2\delta} \end{pmatrix}.$$

Model (3) was also used for the ratoon cycle, with $u_{1i} = 0 \ \forall i$, $i = 1, 2, ..., 39$. The varIdent variance function was used; it is suitable when the data exhibit less variation within a group. For the specification of $\mathbf{R}_i$, $\mathbf{C}_i = \mathbf{I}$ was assumed, and the elements of $\boldsymbol{\Lambda}_i$ are defined by the varIdent variance function; their dependence on time is given by $\lambda_j = \delta_{S_{ij}}$. We set $\delta_{S_{ij}} = 1$ for $\forall$ i and j; then,

 doi:10.20944/preprints202409.0179.v1

$$\boldsymbol{R}_i = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For the plant cycle, the intraclass correlation coefficient ρ_{ICC} was calculated; it indicates the degree of correlation that exists between the observations of the same group. It is calculated using the following equation:

$$\rho_{ICC} = \frac{\sigma_{(u_0,u_1)}}{\sigma_{u_0}\sigma_{u_1}}, \tag{4}$$

where $\sigma_{(u_0,u_1)}$ is the covariance between the random effects of the intercept and slope, and σ_{u_0} and σ_{u_1} are the standard deviations of the random intercept and the random slope, respectively.

For the clustering analysis, two partition-based algorithms, k-means and k-medoids, were utilized [19]; and one density-based algorithm that is robust to outliers, DBSCAN [20].

The evaluation of the clusters identified was done using the silhouette index, the Dunn index, and the connectivity index.
The silhouette index measures the confidence with which an observation is assigned to a cluster, providing information about the cohesion and separation of the clusters. This index varies in the range from -1 to 1, where a value close to 1 indicates that the point has been correctly assigned to the cluster and is well separated from other clusters [21]. Values from 0.71 to 1, a strong structure has been found; from 0.51 to 0.70, a reasonable structure; from 0.26 to 0.50, the structure is weak; and $\leq$0.25, no substantial structure [22]. Negative values have similar interpretation and indicates that the object may be incorrectly assigned and is closer to points from another cluster [23]. The silhouette index is given by

$$I_S = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \tag{5}$$

where n is the total number of objects in the dataset, K is the number of clusters, C_k is the set of objects in cluster k, $a(i)$ is the average distance between object i and all other objects in the same cluster, $b(i)$ is the average distance between object i and all objects in the nearest cluster, and $\max\{a(i), b(i)\}$ is the maximum value of $a(i)$ and $b(i)$.

The Dunn index is a measure of the quality of a partition in terms of the ratio between the minimum distance between cluster centroids and the maximum distance between points and their respective centroids [24]. This index is given by

$$I_D = \frac{\min_{k=1..K}\{d(C_k, C_0)\}}{\max_{x_i\in\Omega}\{d(x_i, c_k)\}}, \tag{6}$$

where K is the number of clusters, $\min_{k=1,\ldots,K}\{d(C_k, C_0)\}$ is the minimum of the distances between the centroids of each cluster C_k and the global center C_0, and $\max_{x_i\in\Omega}\{d(x_i, c_k)\}$ is the maximum distance between each point x_i in the dataset Ω and its respective centroid C_k in its corresponding cluster. The minimum distance is a measure of the internal coherence of the clusters and the maximum distance is a measure of the separation between clusters.

The connectivity index is calculated by constructing a neighbors matrix. Define $nn_{i(j)}$ as the j-th nearest neighbor of observation i, and let $x_{i,nn_{i(j)}}$ be zero if i and $nn_{i(j)}$ are in the same cluster and $1/j$

 doi:10.20944/preprints202409.0179.v1

otherwise. The connectivity index has a value between 0 and ∞ and should be minimized [25]. This index is given by

$$I_C = \sum_{i=1}^{n} \sum_{j=1}^{L} x_{i,nn_{i(j)}}, \tag{7}$$

where n is the total number of objects in the dataset, and L is a parameter that determines the number of neighbors that contribute to the connectivity measure.

3. Results

3.1. Plant Cycle

Table 3 presents the parameter estimates of the fitted model (Model 3), along with their respective 95% confidence interval (CI) for SP, showing both fixed and random effects. The fixed effects include the estimated parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$, both significant, which represent the intercept and slope, respectively. The estimated intercept reflects the average SP value when time is zero, that is, the expected SP at the start of the measurement. The slope indicates the change in SP for each unit change in time; thus, a value of 1.50 signifies that, on average, SP increases by 1.50 units for each additional unit of time. The fitted model for the plant cycle satisfies the assumptions of normality (Kolmogorov-Smirnov, p-value = 0.85), homogeneity of variance (Bartlett, p-value = 0.34), and independence (Box-Pierce, p-value = 0.23). The correlation coefficient (ρ_{ICC}) suggests a negative correlation, indicating a strong inverse relationship between the intercept and the slope. This means that when the intercept is high, the slope tends to decrease, and vice versa. This information is valuable for understanding how different materials will respond and evolve in terms of SP over time. The parameter estimates are presented below.

Table 3. Estimated parameters of fixed and random effects of Model (3) for the plant cycle.

Fixed effects		Random effects			
Parameter	95% CI	Parameter	95% CI	Parameter	95% CI
$\widehat{\beta}_0 = 6.37$	$(5.50, 7.24)$	$\widehat{\sigma}_{u_0} = 2.21$	$(1.56, 3.14)$	$\widehat{\rho}_{ICC} = -0.89$	$(-0.98, -0.50)$
$\widehat{\beta}_1 = 1.50$	$(1.25, 1.75)$	$\widehat{\sigma}_{u_1} = 0.47$	$(0.23, 0.94)$	$\widehat{\sigma}_e = 1.23$	$(1.05, 1.44)$

The estimated parameters for the 39 materials from Las Pilas are shown in Table 4. The variety COLMEX 94-8, which is known to have an Early maturity, has an estimated intercept of 8.97 and an estimated slope of 1.22; similar values are observed for the variety RB 85-5113, which has an estimated intercept of 8.23 and an estimated slope of 1.23. The variety MEX 69-290, which is known to have an Intermediate-Late maturity, has an estimated intercept of 5.67 and an estimated slope of 1.82. From the estimations of all materials, it can be deduced that materials with an Early maturity tend to have a higher intercept. Since both the intercept and the slope are random, their values vary depending on the specific sugarcane material being evaluated. Thus, materials that have a Late maturity tend to have a smaller intercept and a larger slope.

 doi:10.20944/preprints202409.0179.v1

Table 4. Estimated parameters of Model (3) for the 39 materials for the plant cycle .

Material	$\hat{\beta}_0$	$\hat{\beta}_1$	Material	$\hat{\beta}_0$	$\hat{\beta}_1$
COSTA JAL	7.04	1.43	ITSAMEX 07-7259	4.07	2.00
RB 85-5035	7.05	1.34	ITSAMEX 06-6395	9.19	0.98
ITV 92-1424*	8.39	1.21	ITSAMEX 06-4863	5.51	1.73
RB 85-5113*	8.23	1.23	ITSAMEX 07-4954	3.62	1.81
LAICA 92-13	7.11	1.24	ITSAMEX 07-4387	4.68	1.76
CP 72-2086*	7.92	1.28	ITSAMEX 07-121115	3.01	2.14
COLMEX 94-8*	8.97	1.22	ITSAMEX 07-246	3.70	2.00
ATEMEX 99-48	9.79	0.98	ITSAMEX 07-12116	5.45	1.64
ATEMEX 99-1	8.08	1.01	ITSAMEX 07-12113	5.01	1.68
ATEMEX 99-61	9.48	0.94	ITSAMEX 07-2963	5.67	1.59
MEX 70-486	7.97	1.34	ITSAMEX 07-99711	6.43	1.42
TCP 89-3505	3.79	1.85	ITSAMEX 07-9886	6.65	1.41
MEX 80-1521	7.79	1.19	ITSAMEX 07-86810	4.21	1.88
MEX 69-290*	5.67	1.82	ITSAMEX 07-12119	5.04	1.79
ITSAMEX 07-44814	7.80	1.06	ITSAMEX 07-1903	6.78	1.39
ITSAMEX 06-3049	5.60	1.70	ITSAMEX 07-44813	7.56	1.37
ITSAMEX 07-8681	5.92	1.61	ITSAMEX 07-12118	5.52	1.84
ITSAMEX 07-20810	4.21	1.66	CP 85-1382	9.86	0.92
ITSAMEX 07-7501	3.03	2.17	COLMEX 95-27*	7.56	1.31
ITSAMEX 07-1107	5.13	1.56			

[1] $*$ Commercial variety.

For the commercial varieties, the maturity type (Early, Early-Intermediate, Intermediate-Late, and Late) is known for the soil and climate conditions of the region [26]. Meanwhile, for the new materials, the process of identifying the maturity type based on the climate and soil conditions being evaluated is being carried out. Taking this information into account, the materials are grouped.

3.1.1. Clustering

Clustering was performed using the intercept and slope estimates of the model. Since there are four maturity types among the control materials, the number of groups was set to four according to the maturity types of the commercial materials.

The clustering of the materials using the k-means algorithm with $k = 4$ is shown in Figure 1. The variety MEX 69-290, is known to have an intermediate-Late maturity.and belongs to the group in orange which consists of 11 new materials so this group has an Intermediate-Late maturity. The Early group (green) contains ten new materials and The varieties ITV 92-1424, RB 85-5113, CP 72-2086, and COLMEX 95-27 are known to have an Early maturity and belong to the green group formed by other ten new materials, so this is the Early maturity group. The materials shown in brown are classified as having a Late maturity; this group comprises eight new materials. Finally, the Early-Intermediate Group (yellow) contains four new materials and the variety COLMEX 94-8, which is known to have an Early-Intermediate maturity.

 doi:10.20944/preprints202409.0179.v1

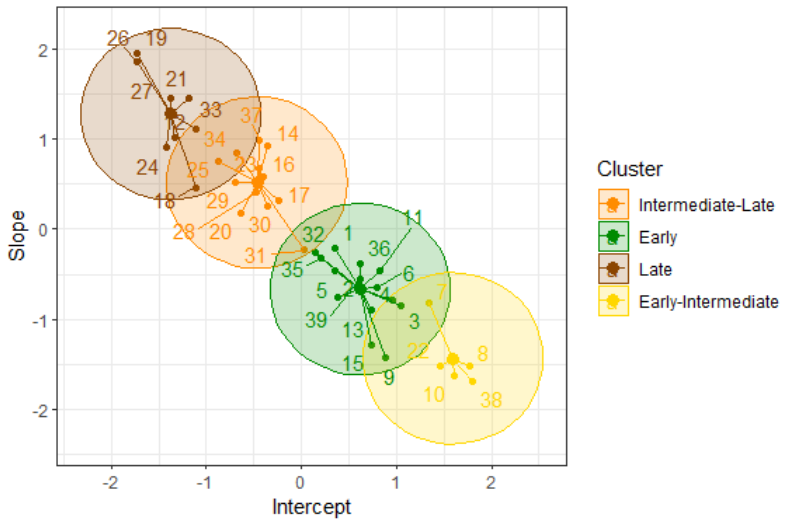

Figure 1. Clustering of the Las Pilas materials in the plant cycle using k-means.

The validation indexes for clustering the materials with k-means are $I_S = 0.63$, $I_D = 0.12$, and $I_C = 13.41$. The value of $I_S = 0.63$ indicates that the grouping of materials shows moderately good cohesion and appropriate separation between the clusters. The value of $I_D = 0.12$ indicates relatively low separation between clusters and a potential for improvement in cohesion within the clusters. Meanwhile, the value of $I_C = 13.41$ indicates good separation between clusters and appropriate compactness within the clusters. Therefore, two out of three indexes indicate that the k-means clustering is good.

The clustering of the materials using the k-medoids algorithm with $k = 4$ is shown in Figure 2. The yellow group contains seven materials classified as having an Early-Intermediate maturity. The green group consists of 12 materials, including the commercial varieties ITV 92-1424, RB 85-5113, CP 72-2086, and COLMEX 94-8, all of which are reported to have an Early maturity. The brown group contains eight materials with a Late maturity. Finally, the orange group includes 11 materials with an Intermediate-Late maturity, as it incorporates the variety MEX 69-290.

 doi:10.20944/preprints202409.0179.v1

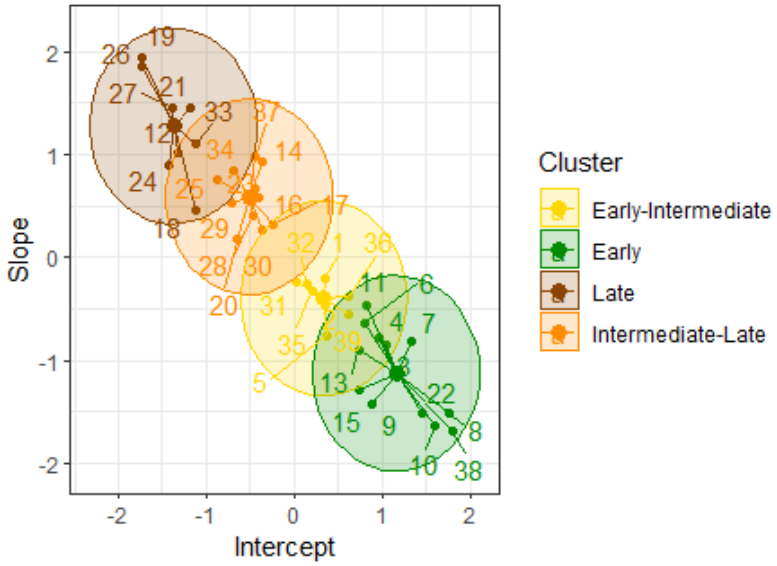

Figure 2. Clustering of the Las Pilas materials in the plant cycle using k-medoids.

The validation indexes for clustering the materials with k-medoids clustering are $I_S = 0.55$, $I_D = 0.12$, and $I_C = 13.85$. The value of $I_S = 0.55$ indicates that the clustering exhibits moderate cohesion and appropriate separation among the groups, but there is still room for improvement in terms of group separation. The value of $I_D = 0.12$ indicates relatively low separation between clusters and a potential for improvement in cohesion within the clusters. The value of $I_C = 13.85$ indicates good separation between the groups and appropriate compactness within the clusters.

The clustering of the materials using the DBSCAN algorithm with $k = 4$ clusters is shown in Figure 3. The Early group (green) consists of 11 new materials and four commercial varieties (ITV 92-1434, RB 85-5113, CP 72-2086, and COLMEX 95-27). The Early-Intermediate group (yellow) includes four materials and the commercial variety COLMEX 94-8. The Intermediate-Late group (orange) comprises 16 new materials, along with the commercial variety MEX 69-290. Lastly, the Late group (brown) consists of two new materials.

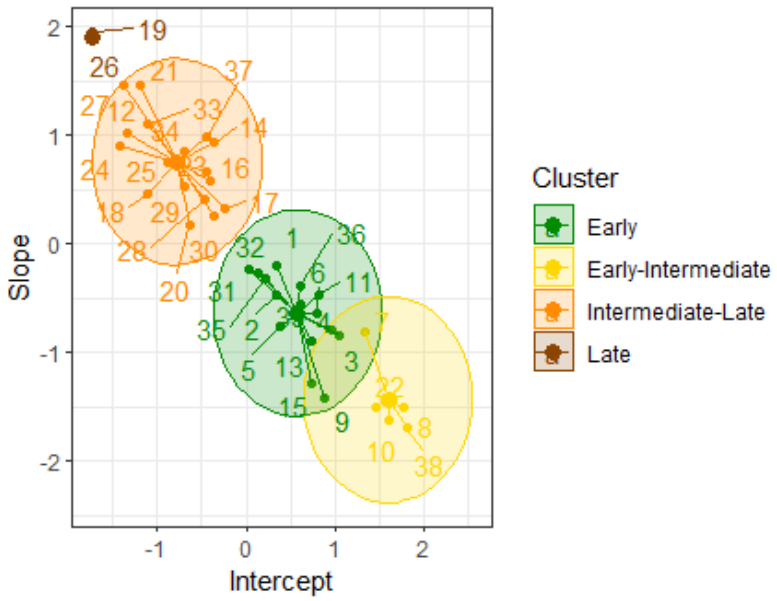

Figure 3. Clustering for the Las Pilas materials in the plant cycle using DBSCAN.

 doi:10.20944/preprints202409.0179.v1

The validation indexes for clustering the materials with DBSCAN are $I_S = 0.64$, $I_D = 0.24$, and $I_C = 12.36$. The values of $I_S = 0.64$ and $I_D = 0.24$ indicate that the clustering exhibits moderately good cohesion and adequate separation between the groups. The value of $I_C = 12.36$ suggests reasonable connectivity between the groups, implying that there is internal connectedness within each group and good relationships between observations in each group. Thus, it can be concluded that the clustering performed with DBSCAN has achieved satisfactory separation and cohesion among the groups, along with suitable connectivity.

Table 5 presents the maturity classification of the materials according to the algorithm used. This table shows the similarities and differences in the classifications made by the algorithms. The k-means and DBSCAN algorithms are more similar to each other than to k-medoids, as they coincide for 32 materials in terms of the maturity type assigned. For k-means and k-medoids, 26 coincidences were found in the maturity type assigned. Lastly, k-medoids and DBSCAN only have 20 coincidences. Thus, the DBSCAN algorithm proved to be the most effective, as the group means were not statistically equal.

 doi:10.20944/preprints202409.0179.v1

Table 5. Comparison of the clustering results for the plant cycle.

ID	Material	k-means	k-medoids	DBSCAN
1	COSTA JAL	Early	Early-Intermediate	Early
2	RB 85-5035	Early	Early-Intermediate	Early
3	ITV 92-1424*	Early	Early	Early
4	RB 85-5113*	Early	Early	Early
5	LAICA 92-13	Early	Early-Intermediate	Early
6	CP 72-2086*	Early	Early	Early
7	COLMEX 94-8*	Early-Intermediate	Early	Early-Intermediate
8	ATEMEX 99-48	Early-Intermediate	Early	Early-Intermediate
9	ATEMEX 99-1	Early	Early	Early
10	ATEMEX 99-61	Early-Intermediate	Early	Early-Intermediate
11	MEX 70-486	Early	Early	Early
12	TCP 89-3505	Late	Late	Intermediate-Late
13	MEX 80-1521	Early	Early	Early
14	MEX 69-290*	Intermediate-Late	Intermediate-Late	Intermediate-Late
15	ITSAMEX 07-44814	Early	Early	Early
16	ITSAMEX 06-3049	Intermediate-Late	Intermediate-Late	Intermediate-Late
17	ITSAMEX 07-8681	Intermediate-Late	Intermediate-Late	Intermediate-Late
18	ITSAMEX 07-20810	Late	Late	Intermediate-Late
19	ITSAMEX 07-7501	Late	Late	Late
20	ITSAMEX 07-1107	Intermediate-Late	Intermediate-Late	Intermediate-Late
21	ITSAMEX 07-7259	Late	Late	Intermediate-Late
22	ITSAMEX 06-6395	Early-Intermediate	Early	Early-Intermediate
23	ITSAMEX 06-4863	Intermediate-Late	Intermediate-Late	Intermediate-Late
24	ITSAMEX 07-4954	Late	Late	Intermediate-Late
25	ITSAMEX 07-4387	Intermediate-Late	Intermediate-Late	Intermediate-Late
26	ITSAMEX 07-121115	Late	Late	Late
27	ITSAMEX 07-246	Late	Late	Intermediate-Late
28	ITSAMEX 07-12116	Intermediate-Late	Intermediate-Late	Intermediate-Late
29	ITSAMEX 07-12113	Intermediate-Late	Intermediate-Late	Intermediate-Late
30	ITSAMEX 07-2963	Intermediate-Late	Intermediate-Late	Intermediate-Late
31	ITSAMEX 07-99711	Intermediate-Late	Early-Intermediate	Early
32	ITSAMEX 07-9886	Early	Early-Intermediate	Early
33	ITSAMEX 07-86810	Late	Late	Intermediate-Late
34	ITSAMEX 07-12119	Intermediate-Late	Intermediate-Late	Intermediate-Late
35	ITSAMEX 07-1903	Early	Early-Intermediate	Early
36	ITSAMEX 07-44813	Early	Early-Intermediate	Early
37	ITSAMEX 07-12118	Intermediate-Late	Intermediate-Late	Intermediate-Late
38	CP 85-1382	Early-Intermediate	Early	Early-Intermediate
39	COLMEX 95-27*	Early	Early-Intermediate	Early

[1] $*$ Commercial variety.

3.2. Ratoon Cycle

For this cycle, incorporating only the random intercept was sufficient to improve the model fit (Model 3), which allowed for capturing the variation in the intercept, indicating notable differences in the initial SP among the different genotypes. Table 6 shows that both the intercept and the slope of the fitted model for the ratoon cycle are significant, with narrow confidence intervals, indicating high precision in the estimates. Additionally, the model satisfies the assumptions of normality (Kolmogorov-Smirnov, p-value = 0.85), homogeneity of variances (Bartlett, p-value = 0.34), and independence (Box-Pierce, p-value = 0.23). The significant variability among the subject intercepts ($\widehat{\sigma}_{u_0} = 1.18$) and the moderate residual variability ($\widehat{\sigma}_e = 1.05$) suggest that the model adequately captures the variation present in the data.

 doi:10.20944/preprints202409.0179.v1

Table 6. Estimated parameters of fixed and random effects of Model (3) for the ratoon cycle.

Fixed effects		Random effects	
Parameter	95% CI	Parameter	95% CI
$\widehat{\beta}_0 = 7.34$	$(6.79, 7.90)$	$\widehat{\sigma}_{u_0} = 1.18$	$(0.90, 1.54)$
$\widehat{\beta}_1 = 1.95$	$(1.25, 1.75)$	$\widehat{\sigma}_e = 1.05$	$(0.93, 1.20)$

The use of growth curve models reveals that the materials exhibit different SP growth curves. These differences are observed among hybrids, varieties, cycles, and even within the same hybrid, which can show distinct behaviors from one cycle to another.

Table 7 displays the estimated parameters for the 39 materials. Similar to the plant cycle, a higher intercept is observed in varieties known to have an Early maturity: ITV 92-1424 had an intercept of 8.30, CP 72-2086 had an intercept of 9.35, and COLMEX 95-27 had an intercept of 9.37. Higher intercepts were also observed for the Early-Intermediate maturity variety: COLMEX 94-8 had an intercept of 9.20. A lower intercept is observed in the MEX 69-290 variety, which had an intercept of 7.16.

Table 7. Estimated parameters for the 39 materials in the ratoon cycle.

Material	$\widehat{\beta}_0$	$\widehat{\beta}_1$	Material	$\widehat{\beta}_0$	$\widehat{\beta}_1$
COSTA JAL	7.00	1.95	ITSAMEX 07-7259	6.78	1.95
RB 85-5035	6.23	1.95	ITSAMEX 06-6395	8.22	1.95
ITV 92-1424*	8.30	1.95	ITSAMEX 06-4863	7.33	1.95
RB 85-5113*	8.30	1.95	ITSAMEX 07-4954	4.48	1.95
LAICA 92-13	6.03	1.95	ITSAMEX 07-4387	5.80	1.95
CP 72-2086*	9.35	1.95	ITSAMEX 07-121115	6.73	1.95
COLMEX 94-8*	9.20	1.95	ITSAMEX 07-246	6.45	1.95
ATEMEX 99-48	9.20	1.95	ITSAMEX 07-12116	6.36	1.95
ATEMEX 99-1	7.40	1.95	ITSAMEX 07-12113	6.51	1.95
ATEMEX 99-61	8.23	1.95	ITSAMEX 07-2963	7.11	1.95
MEX 70-486	8.83	1.95	ITSAMEX 07-99711	7.82	1.95
TCP 89-3505	6.19	1.95	ITSAMEX 07-9886	7.57	1.95
MEX 80-1521	7.37	1.95	ITSAMEX 07-86810	8.03	1.95
MEX 69-290*	7.16	1.95	ITSAMEX 07-12119	7.83	1.95
ITSAMEX 07-44814	7.44	1.95	ITSAMEX 07-1903	8.69	1.95
ITSAMEX 06-3049	6.71	1.95	ITSAMEX 07-44813	6.46	1.95
ITSAMEX 07-8681	6.51	1.95	ITSAMEX 07-12118	7.26	1.95
ITSAMEX 07-20810	6.53	1.95	CP 85-1382	7.87	1.95
ITSAMEX 07-7501	7.00	1.95	COLMEX 95-27*	9.37	1.95
ITSAMEX 07-1107	6.62	1.95			

[1] $*$ Commercial variety.

The estimates of the models for the plant and ratoon cycles show some differences in terms of the estimated intercept. For instance, the COLMEX 95-27 variety had an SP of 7.56 in the plant cycle and an SP of 9.37 in the ratoon cycle. The CP 72-2086 variety had an SP of 7.92 in the plant cycle and an SP of 9.35 in the ratoon cycle. The MEX 69-290 variety had an SP of 5.67 in the plant cycle and an SP of 7.16 in the ratoon cycle. Varieties with smaller variations included ITV 92-1434, with an SP of 8.39 in the plant cycle and an SP of 8.30 in the ratoon cycle, and RB 85-5113, with an SP of 8.23 in the plant cycle and an SP of 8.30 in the ratoon cycle. Finally, the COLMEX 94-8 variety had an SP of 8.97 in the plant cycle and an SP of 9.20 in the ratoon cycle.

 doi:10.20944/preprints202409.0179.v1

3.2.1. Clustering

The clustering of materials using the k-means algorithm is shown in Figure 4. The Early group (green) consists of six new materials and five commercial varieties: ITV 92-1424, RB 85-5113, CP 72-2086, COLMEX 94-8, and COLMEX 95-27. The Early-Intermediate group (yellow) is composed of 14 materials, including the MEX 69-290 variety. Despite being identified as Intermediate-Late in the plant cycle, the MEX 69-290 variety was assigned to the Early-Intermediate cluster in this cycle due to its higher SP. Regarding the absence of the COLMEX 94-8 variety in the Early-Intermediate maturity group, its maturity is determined based on the average SP value that the hybrids in this group have (7.44%). The Late group (brown) contains only the ITSAMEX 07-4954 material. Finally, we have a group of 14 materials that are of Intermediate-Late maturity (orange). Their maturity is inferred from the average value (6.42%) of sucrose that the hybrids exhibit.

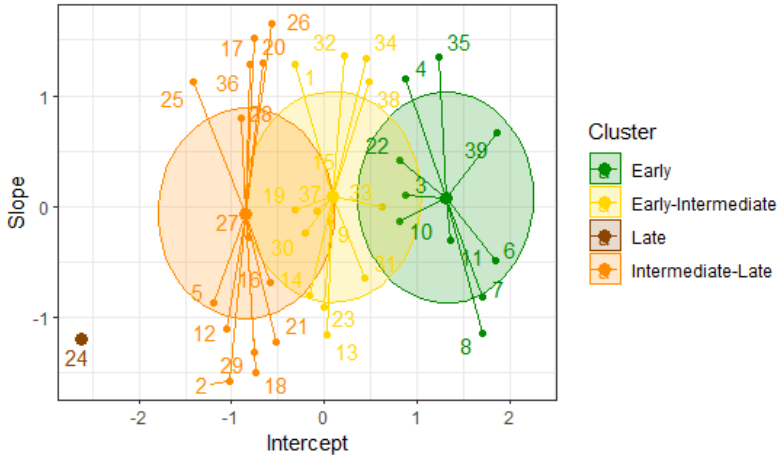

Figure 4. Clustering of the materials in the ratoon cycle using the k-means algorithm.

According to the silhouette index ($I_S = 0.56$), the clustering performed with the k-means algorithm indicates moderate cohesion and satisfactory separation between the clusters. However, it is observed that the separation between groups could be improved, as the Dunn index was low ($I_D = 0.16$). Finally, according to the connectivity index ($I_C = 12.08$), the groups are well-defined and internally connected.

The clustering of materials using the k-medoids algorithm is shown in Figure 5. The Early group (green) is composed of nine materials, including the commercial varieties ITV 92-1434 and RB 85-5113. The Early-Intermediate group (yellow) contains two hybrids and the commercial varieties CP 72-2086, COLMEX 94-8, and COLMEX 95-27. Similarly, the Intermediate-Late group consists of the commercial variety Mex 69-290 and nine hybrids, while the Late group contains 15 hybrids.

 doi:10.20944/preprints202409.0179.v1

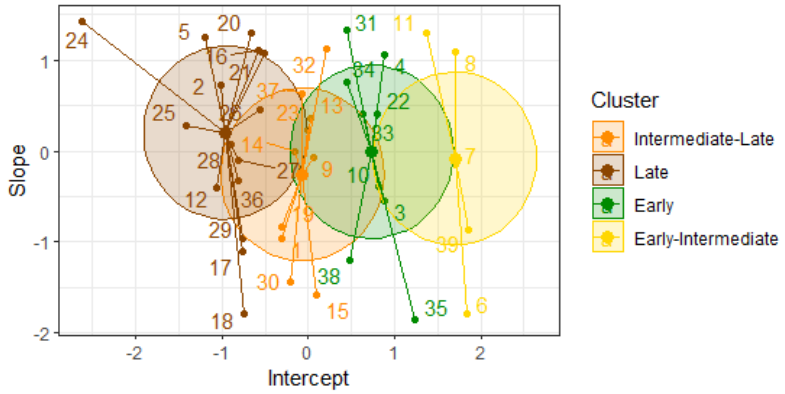

Figure 5. Clustering of the materials in the ratoon cycle using the k-medoids algorithm.

These groups had clustering validation indexes with the following values: $I_S = 0.59$, $I_D = 0.06$ and $I_C = 13.68$. The results indicate that the clustering carried out through the k-medoids algorithm forms groups in which there are similar characteristics among the materials within the same group in terms of β_0 and β_1. The groups show good cohesion and appropriate separation.

Since DBSCAN forms groups based on the data density, these groups can have irregular shapes and sizes. In Figure 6, some groups are more compact, with a low point density (Early-Intermediate and Late groups), while others are more dispersed, with a high point density (Intermediate-Late and Early groups). Thus, the Early maturity (green) group consists of the commercial varieties ITV 92-1434 and RB 85-5113 and six new materials. The Early-Intermediate maturity (yellow) group contains the commercial varieties CP 72-2086, COLMEX 94-8, and COLMEX 95-27 and one hybrid. The Intermediate-Late maturity (orange) group includes 24 new materials and the Mex 69-290 variety. Lastly, the Late group (brown) contains two new materials. Additionally, an outlier material was not assigned to any group. For this reason, the hybrid ITSAMEX 07-4954 was not incorporated into any group for this second cycle.

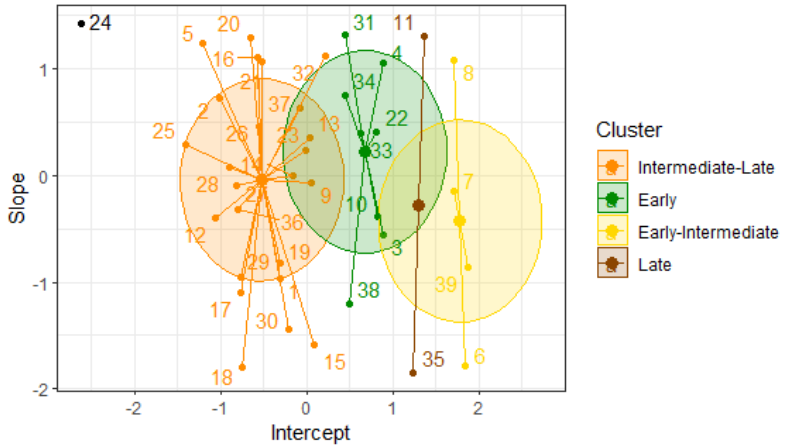

Figure 6. Clustering of the materials in the ratoon cycle using the DBSCAN algorithm.

For this clustering, the validation indices are $I_S = 0.64$, $I_D = 0.14$, and $I_C = 16.11$. These values indicate good cohesion and appropriate separation between the groups. The separation between the groups is reflected in the high value of the Dunn index. These results suggest that the formed groups represent different maturity types of sugarcane and exhibit distinctive characteristics in terms of their density.

 doi:10.20944/preprints202409.0179.v1

Based on the grouping of materials for the k-means algorithm in the plant and ratoon cycles, it was found that the highest sucrose accumulation is observed in the Early-Intermediate and Early groups, respectively. This is because the materials in these groups have a higher average SP. On the other hand, using the k-medoids algorithm revealed that the Early and Early-Intermediate groups accumulated a higher percentage of sucrose. However, with the DBSCAN algorithm, it was determined that the highest SP occurred in the Early-Intermediate group for both cycles. These findings can be attributed to the characteristics and properties of each algorithm, as well as the specificities of each cycle. Therefore, the materials that should be studied in the next phase are COSTA JAL, ATEMEX 99-48, ATEMEX 99-1, ATEMEX 99-61, MEX 70-486, MEX 80-1521, ITSAMEX 07-44814, ITSAMEX 06-6395, ITSAMEX 07-1903, and CP 85-1382. Given that genetic improvement is involved, there is a special interest in selecting varieties with a high sucrose content and early maturation [27].

Table 8 presents a comparison of the clustering algorithms for the ratoon cycle. It can be observed that the k-medoids and DBSCAN algorithms have more coincidences in the assigned maturity type (23 materials); k-means and k-medoids share five assignments, and k-means and DBSCAN have 17 shared assignments. The Intermediate-Late and Early-Intermediate groups have more materials according to the k-means algorithm, while according to the k-medoids algorithm, the Late and Intermediate-Late groups have more assigned materials. Finally, the Intermediate-Late and Early groups have more assigned materials according to the DBSCAN algorithm.

 doi:10.20944/preprints202409.0179.v1

Table 8. Comparison of the algorithms for the ratoon cycle .

ID	Material	k-means	k-medoids	DBSCAN
1	COSTA JAL	Early-Intermediate	Intermediate-Late	Intermediate-Late
2	RB 85-5035	Intermediate-Late	Intermediate-Late	Intermediate-Late
3	ITV 92-1424*	Early	Early	Early
4	RB 85-5113*	Early	Early	Early
5	LAICA 92-13	Intermediate-Late	Late	Intermediate-Late
6	CP 72-2086*	Early	Early-Intermediate	Early-Intermediate
7	COLMEX 94-8*	Early	Early-Intermediate	Early-Intermediate
8	ATEMEX 99-48	Early	Early-Intermediate	Early-Intermediate
9	ATEMEX 99-1	Early-Intermediate	Intermediate-Late	Intermediate-Late
10	ATEMEX 99-61	Early	Early	Early
11	MEX 70-486	Early	Early-Intermediate	Late
12	TCP 89-3505	Intermediate-Late	Late	Intermediate-Late
13	MEX 80-1521	Early-Intermediate	Intermediate-Late	Intermediate-Late
14	MEX 69-290*	Early-Intermediate	Intermediate-Late	Intermediate-Late
15	ITSAMEX 07-44814	Early-Intermediate	Intermediate-Late	Intermediate-Late
16	ITSAMEX 06-3049	Intermediate-Late	Late	Intermediate-Late
17	ITSAMEX 07-8681	Intermediate-Late	Late	Intermediate-Late
18	ITSAMEX 07-20810	Intermediate-Late	Late	Intermediate-Late
19	ITSAMEX 07-7501	Early-Intermediate	Intermediate-Late	Intermediate-Late
20	ITSAMEX 07-1107	Intermediate-Late	Late	Intermediate-Late
21	ITSAMEX 07-7259	Intermediate-Late	Late	Intermediate-Late
22	ITSAMEX 06-6395	Early	Early	Early
23	ITSAMEX 06-4863	Early-Intermediate	Intermediate-Late	Intermediate-Late
24	ITSAMEX 07-4954	Late	Late	Null
25	ITSAMEX 07-4387	Intermediate-Late	Late	Intermediate-Late
26	ITSAMEX 07-121115	Intermediate-Late	Late	Intermediate-Late
27	ITSAMEX 07-246	Intermediate-Late	Late	Intermediate-Late
28	ITSAMEX 07-12116	Intermediate-Late	Late	Intermediate-Late
29	ITSAMEX 07-12113	Intermediate-Late	Late	Intermediate-Late
30	ITSAMEX 07-2963	Early-Intermediate	Intermediate-Late	Intermediate-Late
31	ITSAMEX 07-99711	Early-Intermediate	Early	Early
32	ITSAMEX 07-9886	Early-Intermediate	Intermediate-Late	Intermediate-Late
33	ITSAMEX 07-86810	Early-Intermediate	Early	Early
34	ITSAMEX 07-12119	Early-Intermediate	Early	Early
35	ITSAMEX 07-1903	Early	Early-Intermediate	Late
36	ITSAMEX 07-44813	Intermediate-Late	Late	Intermediate-Late
37	ITSAMEX 07-12118	Early-Intermediate	Intermediate-Late	Intermediate-Late
38	CP 85-1382	Early-Intermediate	Early	Early
39	COLMEX 95-27*	Early	Early-Intermediate	Early-Intermediate

[1] ∗ Commercial variety.

Based on these results, it can be deduced that the DBSCAN algorithm remained consistent in the identification and characterization of materials, which means that it tends to generate more similar or stable results across different runs or configurations. Furthermore, it is widely recognized as one of the most powerful algorithms in the field of clustering and is notable for its density-based clustering capabilities [28].

4. Discussion

Improving sugarcane varieties with high yield and higher sucrose content is essential for the sugar industry. Previous studies, such as those by [15] and [29], have emphasized the importance of these goals. The present research analyzes how sucrose accumulation varies among different hybrids and varieties, and how this variability is observable in both the plant cycle and the ratoon cycle.

 doi:10.20944/preprints202409.0179.v1

Variations in sucrose accumulation among different materials have been observed to be due to various factors, including genetic factors, soil conditions, climate, pests, and agronomic management. [30] and [31] have also identified these factors as critical for the growth and development of sugarcane, as well as for sucrose accumulation and the maturity processes.

A significant finding of this research is that the highest sucrose content occurs in the ratoon cycle compared to the plant cycle. This indicates that management practices and the timing of harvest are crucial for maximizing sucrose content. [32] highlighted the importance of the maturity stage for sucrose storage, and studies such as those by [33] and [34] mention that processing the cane before its optimal maturity can significantly reduce the sucrose content.

On the other hand, [35] obtained higher sucrose yields during the plant cycle when evaluating three levels of fertilization in a vertisol soil. In the plant, ratoon, and second ratoon cycles at intervals of 18, 12, and 12 months, respectively, these authors observed lower yields in the ratoon and second ratoon cycles compared to the plant cycle. This variability indicates that specific agronomic management practices and soil conditions play a crucial role in optimizing yield.

These findings underscore the importance of adapting agronomic practices and selecting specific varieties to maximize yield in different environments. Genetic variability and environmental conditions must be carefully considered to optimize sucrose production. Selecting varieties that are well-suited to local conditions can significantly improve sucrose yields.

Regarding the validation of clustering with the silhouette index, values greater than 0.5 were obtained for all three algorithms in both cycles. These results differ from those obtained by [11] when evaluating the silhouette index in clustering sugarcane genotypes with k-means; these authors obtained a value of 0.34, which is lower than the values obtained in this study. The difference could be attributed to the sugarcane materials used in their research, meaning there may be variations in their characteristics that influence the clustering algorithms' ability to find clear patterns. Additionally, the SP sampling in this study was limited to four events, and the inherent crop conditions could also impact the results. Similarly, variations in the parameters used in the k-means algorithm, such as the number of clusters or the distance criterion, may also affect clustering outcomes.

Future research could focus on expanding SP sampling to more events and different cultivation conditions to gain a more detailed and robust understanding of the variations in sucrose accumulation. Moreover, the implementation of new technologies and analytical methods could improve the precision and efficiency of clustering studies and variety selection.

5. Conclusions

The sucrose accumulation curves of 33 new materials were studied using growth curve models. This enabled the analysis of the material variability over time, the variation in the sucrose yield among the materials, and the behavior of the commercial varieties. Subsequently, the materials were identified and characterized based on the maturities that they exhibited.

In this way, hybrids with higher sucrose accumulations were identified based on the evaluated time period. These hybrids are primarily in the Early-Intermediate and Early groups. The materials proposed for further study in the subsequent phase include COSTA JAL, ATEMEX 99-48, ATEMEX 99-1, ATEMEX 99-61, MEX 70-486, MEX 80-1521, ITSAMEX 07-44814, ITSAMEX 06-6395, ITSAMEX 07-1903, and CP 85-1382, as they demonstrated better adaptation to the established conditions.

The characterization of the 33 new materials in the adaptability testing phase could be valuable for plant breeders and geneticists searching for new commercial sugarcane varieties. The methodology employed here could serve as a protocol for studying and characterizing sugarcane genotypes or identifying new commercial varieties in the advanced stages of selection. Additionally, it can aid in selecting materials for further study. The accurate identification and characterization of materials offer the sugar industry insights for improved harvest planning, effective agricultural management, and informed decision-making regarding sugarcane processing.

 doi:10.20944/preprints202409.0179.v1

Author Contributions: Conceptualization, M.G.M. and C.D.C.M.; methodology, M.G.M., C.D.C.M. and F.G.J.; software, C.D.C.M., M.G.M. and F.G.J.; validation, M.G.M., C.D.C.M., J.C.G.P., E.P.A. and R.R.C.; formal analysis, M.G.M., C.D.C.M. and J.C.G.P.; investigation, J.C.G.P.; resources, J.C.G.P.; data curation, C.D.C.M. and J.C.G.P.; writing—original draft preparation, C.D.C.M., M.G.M. and F.G.J.; writing—review and editing, M.G.M., C.D.C.M., F.G.J., J.C.G.P. and R.R.C.; visualization, C.D.C.M., M.G.M. and F.G.J; supervision, M.G.M., J.C.G.P., J.T.S. and F.G.J.; project administration, J.C.G.P.; funding acquisition, J.C.G.P. All authors have read and agreed to the published version of the manuscript.

Funding: The project was funded by Fundación Produce de Jalisco, while the land and technical monitoring of the research were carried out with the help of the Sugarcane Production and Quality Committees of Ingenio Melchor Ocampo and INIFAP.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study can be requested from the corresponding author. They are not publicly available because the database was provided by third parties.

Acknowledgments: To the technical field department of Ingenio Melchor Ocampo S.A de C.V., to Engineer Filemón Zavalza García, and to the sugarcane production and quality committee of the mentioned mill. Special thanks to Gabriel Ricardo Blackaller Ayala and Jesús Zúniga Mendoza.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SP	Sucrose Percentage
PP	Polarization Percentage
INIFAP	National Institute of Forestry, Agriculture, and Livestock Research
I_S	Silhouette Index
I_D	Dunn Index
I_C	Connectivity Index

References

1. Alejandre Rosas, J.A.; Galindo Tovar, M.E.; Lee Espinosa, H.E.; Alvarado Gómez, O.G. Variabilidad genética en 22 variedades híbridas de caña de azúcar (Saccharum spp. Híbrido). *Phyton (Buenos Aires)* **2010**, *79*(1), 87–94. Available online: https://revistaphyton.fund-romuloraggio.org.ar/vol79/Alejandre-Rosas.pdf.
2. Arntzen, C.J.; Ritter, E.M. *Encyclopedia of agricultural science Volume 2: EL.*; Academic Press: 1994. Available online: https://www.cabidigitallibrary.org/doi/full/10.5555/19951403738.
3. CONADESUCA. *Nota informativa sobre innovaciones en materia de productividad del sector. Nuevas variedades de caña de azúcar*. Sagarpa: Ciudad de México, 2016. Available online: https://www.gob.mx/cms/uploads/attachment/file/136406/NotaNuevasVariedadesd_Cana_deAzucar.compressed.pdf.
4. Que, Y.; Wu, Q.; Zhang, H.; Luo, J.; Zhang, Y. Developing new sugarcane varieties suitable for mechanized production in China: principles, strategies and prospects. *Frontiers in Plant Science* **2024**, *14*, 1337144. https://doi.org/10.3389/fpls.2023.1337144.
5. Sentíes-Herrera, H. E.; Valdez-Balero, A.; Loyo-Joachin, R.; Gómez-Merino, F. C. Fases experimentales en el mejoramiento genético de la caña de azúcar (*Saccharum spp.*) en México. *AgroProductividad* **2017**, *10*, 93–99. Available online: https://go.gale.com/ps/i.do?p=IFME&u=anon~3406d702&id=GALE|A530914354&v=2.1&it=r&sid=googleScholar&asid=76e56977.
6. García-Preciado, J. C. Evaluación de variables de calidad en híbridos de *Saccharum spp.* en diferentes ambientes agroecológicos de Jalisco, México. *AgroProductividad* **2017**, *10*, 11. Available online: https://www.revista-agroproductividad.org/index.php/agroproductividad/article/view/55.
7. Bastidas, L.; Rea, R.; Sousa Vieira, O. de; Hernández, E.; Briceño, R. Análisis de variables agronómicas en cultivares de caña de azúcar con fines azucareros, paneleros y forrajeros. *Bioagro* **2012**, *24*, 135–142. Available online: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-33612012000200008&nrm=iso.
8. Rodríguez Gross, R.; Puchades Izaguirre, Y.; Abiche Maceo, W.; Rill Martínez, S.; Suarez, H. J.; Salmón Cuspineda, Y.; Gálvez, G. Estudio del rendimiento y modelación del período de madurez en nuevos

 doi:10.20944/preprints202409.0179.v1

cultivares de caña de azúcar. *Cultivos Tropicales* **2015**, *36*, 134–143. Available online: http://scielo.sld.cu/scielo.php?pid=S0258-59362015000400019&script=sci_arttext.

9. Hernández, O.L.; García, S.S.; Nataren, E.H.; Espinoza, L.C.L.; Oliva, A.C.; Sanchez, S.C.; Romero, E.R.; Zossi, S. La espectroscopía de infrarrojo cercano (NIRS) en el seguimiento de la madurez del cultivo de la caña de azúcar (Saccharum spp.). *Agro Productividad* **2019**, *12*(7). https://doi.org/10.32854/agrop.v0i0.1477.
10. Larrahondo, J. E.; Cassalett, C.; Torres, J.; Issacs, C. El cultivo de la caña de azúcar en la zona azucarera de Colombia. In *Calidad de caña*; Cassalett, C., Torres, J., Issacs, C., Eds.; Cenicaña: Cali, 1995; pp. 337–354. Available online: https://www.cenicana.org/pdf_privado/documentos_no_seriados/libro_el_cultivo_cana/libro_p3-394.pdf.
11. Ostengo, S.; Rueda Calderón, M. A.; Bruno, C.; Cuenya, M. I.; Balzarini, M. A protocol for identifying characteristic sucrose accumulation curves of sugarcane genotypes (*Saccharum spp.*). *Sugar Tech* **2021**, *23*, 519–523. https://doi.org/10.1007/s12355-020-00926-8.
12. West, B. T.; Welch, K. B.; Galecki, A. T. *Linear Mixed Models: A Practical Guide Using Statistical Software*; Chapman and Hall/CRC: Boca Raton, FL, 2006; pp. 9–28. https://doi.org/10.1201/b17198.
13. Balaguera Carrasquilla, N. P. *Modelo de curvas de crecimiento: Modelos lineales mixtos vs Datos funcionales*. Ph.D. Thesis, Universidad Santo Tomás, 2020. Available online: http://hdl.handle.net/11634/22477.
14. Gilbert, R. A.; Shine Jr., J. M.; Miller, J. D.; Rice, R. W. Sucrose accumulation and harvest schedule recommendations for CP sugarcane cultivars. *Crop Management* **2004**, *3*, 1–7. https://doi.org/10.1094/CM-2004-0402-01-RS.
15. Zhao, Y.; Liu, J.; Huang, H.; Zan, F.; Zhao, P.; Zhao, J.; Deng, J.; Wu, C. Genetic improvement of sugarcane (Saccharum spp.) contributed to high sucrose content in China based on an analysis of newly developed varieties. *Agriculture* **2022**, *12*(11), 1789. https://doi.org/10.3390/agriculture12111789.
16. McCulloch, C. E.; Searle, S. R. *Generalized, Linear, and Mixed Models*; Wiley-Interscience: Hoboken, NJ, 2001. https://doi.org/10.1002/0471722073.
17. Gałecki, A.; Burzykowski, T. *Linear mixed-effects models using R*; Springer: New York Heidelberg Dordrecht London, 2013; pp. 245–273. https://doi.org/10.1007/978-1-4614-3900-4.
18. Aguirre-Calderón, O.A. ¿Cómo corregir la heterocedasticidad y autocorrelación de residuales en modelos de ahusamiento y crecimiento en altura? *Revista Mexicana de Ciencias Forestales* **2018**, *9*, 49. https://doi.org/10.29298/rmcf.v9i49.151.
19. Zhang, C.; Huang, W.; Niu, T.; Liu, Z.; Li, G.; Cao, D. Review of Clustering Technology and Its Application in Coordinating Vehicle Subsystems. *Automotive Innovation* **2023**, 1–27. https://doi.org/10.1007/s42154-022-00205-0.
20. Khan, M. M. R.; Siddique, M. A. B.; Arif, R. B.; Oishe, M. R. ADBSCAN: Adaptive density-based spatial clustering of applications with noise for identifying clusters with varying densities. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*; IEEE: Dhaka, Bangladesh, 2018; pp. 107–111. https://doi.org/10.1109/CEEICT.2018.8628138.
21. Ben Ncir, C.-E.; Hamza, A.; Bouaguel, W. Parallel and scalable Dunn Index for the validation of big data clusters. *Parallel Computing* **2021**, *102*, 102751. ISSN: 0167-8191. https://doi.org/10.1016/j.parco.2021.102751.
22. Kaufman, L.; Rousseeuw, P.J. *Finding groups in data: an introduction to cluster analysis*; John Wiley & Sons: New Jersey, 2009. Available online: https://books.google.es/books?hl=es&lr=&id=YeFQHiikNo0C&oi=fnd&pg=PR11&dq=Finding+groups+in+data:+an+introduction+to+cluster+analysis&ots=5CofD1KFAE&sig=CaZq_JCe5uV8RB_ymUHSuad8bKQ#v=onepage&q=Finding%20groups%20in%20data:%20an%20introduction%20to%20cluster%20analysis&f=false.
23. Videla, M. E.; Bruno, C. Validación de agrupamientos para representar estructura genética poblacional. *Agriscientia* **2022**, *39*, 1–10. https://doi.org/10.31047/1668.298x.v39.n1.34015.
24. Luna-Romera, J. M.; del Mar Martinez-Ballesteros, M.; Garcia-Gutierrez, J.; Riquelme-Santos, J. C. An approach to silhouette and dunn clustering indices applied to big data in spark. In *Advances in Artificial Intelligence: 17th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2016, Salamanca, Spain, September 14-16, 2016. Proceedings 17*; Springer: Cham, 2016; pp. 160–169. https://doi.org/10.1007/978-3-319-44636-3_15.
25. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R package for cluster validation. *Journal of Statistical Software* **2008**, *25*, 1–22. https://doi.org/10.18637/jss.v025.i04.

 doi:10.20944/preprints202409.0179.v1

26. Gómez-Merino, F. C. *Manual para la identificación varietal de caña de azúcar*; Colegio de Postgraduados: Texcoco, Estado de México, 2015. Available online: https://www.researchgate.net/profile/Fernando-Gomez-Merino/publication/271647569_Manual_para_la_Identificacion_Varietal_de_Cana_de_Azucar/links/54f483720cf2eed5d734bf55/Manual-para-la-Identificacion-Varietal-de-Cana-de-Azucar.pdf.
27. Mendoza Batista, Y.; Cruz Sarmiento, R.; Vaillant Cáceres, Y.; Luis Martínez, O.; Céspedes Argota, M. Comportamiento de los cultivares de caña de azúcar C97-445 y C95-416 en localidades de la provincia Holguín. *Centro Agrícola* **2019**, *46*(1), 49–53. Available online: http://scielo.sld.cu/scielo.php?pid=S0253-57852019000100049&script=sci_arttext&tlng=en.
28. Mei, H.; Mao, L.; Zhang, Y.; Chen, M. BDT-ADBSCAN: Adaptive Density-Based Spatial Clustering of Applications with Noise Based on Bayesian Decision Theory for Identifying Clusters with Multi-Densities. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2022; pp. 1510–1516. https://doi.org/10.1109/ITAIC54216.2022.9836545.
29. Cheavegatti-Gianotto, A.; De Abreu, H.M.C.; Arruda, P.; Bespalhok Filho, J.C.; Burnquist, W.L.; Creste, S.; di Ciero, L.; Ferro, J.A.; de Oliveira Figueira, A.V.; de Sousa Filgueiras, T.; et al. Sugarcane (Saccharum X officinarum): a reference study for the regulation of genetically modified cultivars in Brazil. *Tropical plant biology* **2011**, *4*, 62–89. https://doi.org/10.1007/s12042-011-9068-3.
30. Marcano, M.; Manrique, U.; Garcia, M.; Salcedo, F. Prueba de ocho variedades de caña de azúcar (Saccharum sp.) bajo condiciones de secano en un suelo de sabana del estado Monagas, Venezuela. *Revista Científica UDO Agrícola* **2005**, *5*(1), 54–61. Available online: https://dialnet.unirioja.es/servlet/articulo?codigo=2221593.
31. Vasantha, S.; Kumar, R.A.; Tayade, A.S.; Krishnapriya, V.; Ram, B.; Solomon, S. Physiology of sucrose productivity and implications of ripeners in sugarcane. *Sugar Tech* **2021**, 1–17. https://doi.org/10.1007/s12355-021-01062-7.
32. Jackson, P.A. Breeding for improved sugar content in sugarcane. *Field Crops Research* **2005**, *92*(2-3), 277–290. https://doi.org/10.1016/j.fcr.2005.01.024.
33. Delgado Mora, I.; Jorge Suarez, H.; Vera, A.; Cornide Hernández, M.T.; Díaz Mujica, F.R.; Gómez Pérez, J.R.; Suárez Sanchez, O.; Puchades Isaguirre, Y. Influencia de la edad y cultivar de caña de azúcar en el momento de la cosecha. *Centro Agrícola* **2016**, *43*(2), 59–65. Available online: http://scielo.sld.cu/scielo.php?pid=S0253-57852016000200008&script=sci_arttext.
34. Espinoza, J. G. Maduración de la caña de azúcar y floración de la caña de azúcar y su manejo. *Cengicaña* **2012**, 262–281. Available online: https://cengicana.org/files/20150828053619432.pdf.
35. García, S.S.; Escobar, R.N.; Alanis, L.B. Determinación de la dosis óptima económica de fertilización en caña de azúcar. *Terra latinoamericana* **2003**, *21*(2), 267–272. Available online: http://www.redalyc.org/articulo.oa?id=57315595012.