Cultivos Tropicales, Vol. 43, No. 3, July-September 2022, e-ISSN: 1819-4087, p-ISSN: 0258-5936, https://ediciones.inca.edu.cu

Cu-ID: https://cu-id.com/2050/v43n3e07

Original article

Big data management for sugarcane breeding program

Gestión de macrodatos para el programa de mejora genética de la caña de azúcar

Reynaldo Rodríguez-Gross[1*], Yaquelin Puchades-Isaguirre[1], Wilfre Aiche-Maceo[1], Héctor García-Pérez[2]

[1]Estación Territorial de Investigaciones de la Caña de Azúcar Oriente Sur. Cuba

[2]Instituto de Investigaciones de la Caña de Azúcar (INICA). Cuba

ABSTRACT: The aim of this work was to design and establish a model of macro data management to facilitate decision-making in the Sugarcane Breeding Program in Cuba and increase its efficiency. For this, available sources of information from the selection process from the period 2000 to 2017 and those coming from the agroproductive response of cultivars in production areas were used. A model was designed and applied that includes components such as: infrastructure, collection, validation, storage, processing, analysis and visualization. A case study of the C86-12 x CP70-1133 cross was conducted to characterize its selection background. As a result, the big data approach made it possible to obtain a compilation of primary data and selection results, estimate the genetic value of parents and crosses, classify the crosses and facilitate decision-making in the sugarcane breeding program in Cuba to obtain new commercial cultivars. Its application in the case study guaranteed access to all available information on the subject and recommending its best management.

Key words: hybridization, ICT, statistical methods.

RESUMEN: El objetivo de este trabajo fue diseñar y establecer un modelo de gestión de macrodatos para facilitar la toma de decisiones en el Programa de Mejoramiento de la caña de azúcar en Cuba e incrementar su eficiencia. Para esto se utilizaron las fuentes de información disponibles del proceso de selección del período 2000 al 2017 y las procedentes de la respuesta agroproductiva de los cultivares en áreas de producción. Se diseñó y aplicó un modelo que incluye los componentes: infraestructura, colección, validación, almacenamiento, procesamiento, análisis y visualización. Se realizó un estudio de caso del cruce C86-12 x CP70-1133 para caracterizar sus antecedentes de selección. Como resultado, el enfoque *big data* permitió obtener una compilación de datos primarios y resultados de selección, estimar el valor genético de progenitores y cruces, clasificar los cruzamientos y facilitar la toma de decisiones en el programa de mejoramiento de la caña de azúcar en Cuba para la obtención de nuevos cultivares comerciales. Su aplicación en el caso de estudio garantizó acceder a toda la información disponible al respecto y recomendar su mejor manejo.

Palabras clave: hibridación, TIC, métodos estadísticos.

INTRODUCTION

The use of big data techniques and methods in agriculture is a great opportunity for the use of technologies depending on the investment and the perception of their additional value within the agri-food sector (1,2). The *big data* paradigm is recent and plays an important role in improving the efficiency of the entire supply chain and mitigating food safety issues (3).

Big data applications in agriculture are not strictly concerned with the primary production of large data sets, but base management tasks on interfacing automated forms of collection and storage with pattern recognition functions (4). While the *big data* approach takes into account variability in the field, it requires real-time reconfiguration tools to streamline decision-making, and generally includes intelligent assistance in the implementation, maintenance, and use of the technology (5).

*Author for correspondence: reynaldo.rodriguez@inicasc.azcuba.cu

Received: 20/01/2021

Accepted: 25/07/2021

Sugarcane breeding programs follow four key steps: (i) the generation of a large progeny population from specific crosses, (ii) the evaluation of those progeny at different stages, (iii) clone selection with superior characteristics, and (iv) recombination of the selected clones to close the cycle, or to initiate a new one. This is a relatively long process (never less than 10 years), which consumes resources and generates a large amount of information that must be summarized in different forms (6,7).

The Sugarcane Research Institute of Cuba (INICA) develops a genetic breeding program to respond to the breeding of cultivars (8). This implies that it annually handles an extensive volume of data from a large population of clones and cultivars in different selection stages.

To capture, store and process the information obtained in the Sugarcane breeding Program in Cuba, a computer program (SASEL) is available to improve the efficiency of the whole process (9). In addition to the above, there is information coming from the response of cultivars in production conditions, both yield and reaction to pests, which is not taken into consideration, in a dynamic way, to correct the sugarcane crossing program.

The aim of this work was to design and establish a management model with a macro data approach assisted by the SASEL interface to facilitate decision making in the sugarcane breeding program in Cuba and increase its efficiency.

MATERIALS AND METHODS

For the realization of the model of macro data management of the sugarcane-breeding program, the development of its components was taken into account, such as:

Infrastructure and digital security: Breeding Program with a network of test environments (eleven), which are representative of the main conditions in which sugarcane is grown in the country (Figure 1). It is staffed and maintained by seven provincial experimental stations and six extension and agricultural services groups, all of which have the necessary personnel (researchers, technicians and specialists) and equipment (greenhouses, molecular biology, diagnostic, seed, physiology and sugar analysis laboratories). It also has a center exclusively for generating variability through hybridization. This infrastructure ensures the collection of data from both new cultivar studies and those planted in sugarcane production areas.

Validation of information: The entire process of genetic breeding (including generation of variability, selection, pest resistance tests, extension in commercial areas) is governed by Standards and Procedures (10), which allow standardizing procedures and structuring the databases of the information generated, auditable by the program management, at the same time that it is validated.

Collection, storage and organization of data: The data have a physical record in the file or protocol of each trial. A computer application called SASEL is available to facilitate its digitalization and management (9). This application stratifies the information at the provincial level by year of selection or series, combination, stage and strain, and for the national analysis, it adds the province.

Tools for analysis, modeling and visualization: Genetic-statistical parameters such as mean, standard deviation, variance and differential selection are determined for the selection databases, either from a particular series or from compilation in time and space (several series, stages and localities).

In addition to the genetic-statistical parameters, the individual data of crosses of a series or compiled from several series were used to determine an index of simultaneous selection in the estimation of the genetic value (EVG) of crosses and parents (11). With the EVG values, a cross evaluation methodology was developed to classify crosses from highly proven to highly discard (12). These mathematical tools of multivariate analysis were used as a whole to process the chain of big data generated in the breeding process.

For the data visualization and their analysis, the default reports of SASEL interface were used, in addition to the customized queries that can be performed. Moreover, the database relationship model of the SASEL software platform was restructured to a new big data approach that included relational, semi-relational and non-relational databases:

- Compilation of selection data of primary field from the first and second clonal stage by series and territories.

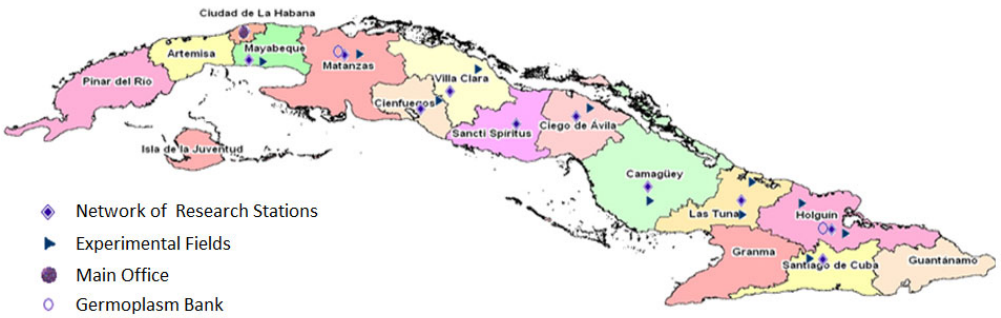


Figure 1. Experimental network for the sugarcane breeding program in Cuba

- Compilation of field selection results of the first and second clonal stage by series and territories.
- Estimation of the genetic value of parents and crosses and classification.

Case study. Selection results in the first and second clonal propagation.

As an example of the *Big-Data* approach, a chain of data resulting from the selection process in clonal stages I and II of the current selection scheme was processed. It is corresponding to the progenies obtained in the period 2000-2007, from the biparental cross between the female and male parents respectively C86-12 and CP70-1133, evaluated in the test sites or selection environments corresponding to Villa Clara, Holguin and Santiago de Cuba provinces.

Data universe included 160 806 clones (It does not quite understand the 1738 crosses, unless the participation of the parents C86-12 and CP70-1133 was taken into account within these crosses) processed with the aid of SASEL application that allowed obtaining the summaries of the selection by series and territories.

RESULTS AND DISCUSSION

The big data management scheme for the sugarcane breeding program in Cuba takes into account the information generated in the selection process of new cultivars, as well as the evaluations carried out in commercial production areas (Figure 2). This design was based on the integration of relational, semi-relational and non-relational databases.

The primary information is stored in different databases such as crossbreeding program, evaluations, and field selection and resistance tests. All these determinations are obtained by series or years, breeding site and stages of the selection scheme.

The evaluations of the agro-productive response of cultivars in production areas and percentages of occupied areas, commercial validation tests of new cultivars and phytosanitary surveys are also part of the *big data* scheme. These databases have between them and the breeding program a relationship that is not always relational or structured; that is, they can be provided in different formats and from different sources.

Big data management allows data to be stored, organized and validated, and then analyzed, modeled and visualized. The core of analysis and processing (Figure 2) is supported by the application of statistical analysis, time and space series analysis, mathematical models and genetic value estimation (11), as well as cross and parent classification methodologies (12). With the results obtained in these analyses, it is facilitated the decision making in the sugarcane breeding program in Cuba, which brings an increase in its efficiency and in cultivar recommendation composition in commercial production areas.

On the other hand, a new option of SASEL interface was obtained for the management and visualization of data in a *big data* environment (Figure 3). It is important to highlight the four fundamental elements that make it up. Box a1 is composed of the databases with the record of parents and crosses to be processed. This information can come from the crossing or hybridization list, the file of the series in use or manually entered data.

Box a2 and a3 distinguish the source or database from which the information on the selected parents and crosses is obtained, and whether this information is primary or previously processed with the application of statistical analysis and mathematical models. The databases that are integrated in this option are: (i) seedling stage; (ii) clonal batches stage 1 and 2; (iii) replicated studies; (iv) Genetic value of crosses; (v) commercial varieties; and (vii) resistance tests, which include phytosanitary surveys.

Box a4 reveals the type of output and the scope of information consulted. It is possible to perform individual analyses by parent, series, territory or stage, and comprehensive summaries can be obtained. It is in the latter that *big data* approach is best appreciated, as the

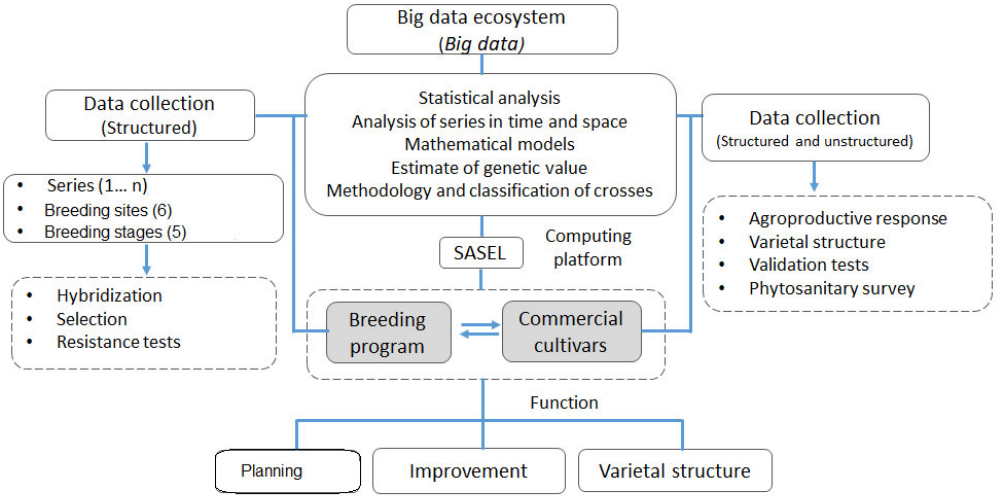
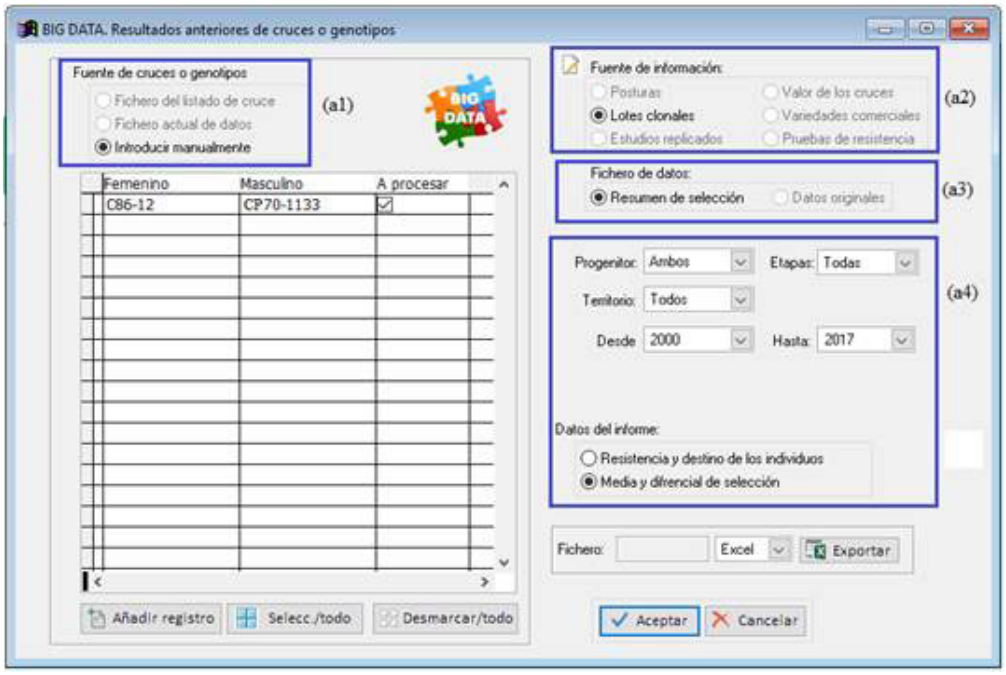


Figure 2. Big data management scheme for the sugarcane genetic breeding program in Cuba



(a1- parent record, a2- information source, a3- data collected or processed, a4- report output)

Figure 3. Interface for handling and visualization of macrodata

information is displayed in summary form, which facilitates understanding and thus decision making.

This interconnection of databases has made it possible to obtain a compilation of results on the number of crosses evaluated by the genetic breeding program, their frequency and results. A mathematical model was also obtained to estimate the genetic value of parents and crosses (11), and it constitutes the basis of the methodology for the classification of crosses (12). This last result takes into consideration the genotype-environment interaction, which undoubtedly constitutes a valuable tool to increase the efficiency of the breeding program.

The incorporation of new data bases from evaluations in commercial production areas offers the possibility of new outlets to improve the recommendations for cultivar composition and management.

Case study. Background of selection results and genetic value in clonal stages 1 and 2

The query performed on the selection background in the databases of clonal batches 1 and 2 of the C86-12 x CP70-1133 cross showed their participation results in the study period of the series from 2000 to 2017 (Table 1). This output is the product of the collection, processing, analysis and interconnection of six data chains:

(a1) Data collected in the selection evaluations of clonal stages 1 and 2. In the study period, only seven series were found where the C86-12 x CP70-1133 cross participated, representing data management in two stages, three provinces or territories (Villa Clara, Holguín and Santiago de Cuba) and three years or series (2008, 2009 and 2011).

a2). It shows the reaction in state resistance tests for brown rust, charcoal, leaf scald and sugarcane virus pests of C86-12 and CP70-1133 cultivars used as female and male parents, respectively. The reaction categories to these four diseases are summarized as susceptible (S), Intermediate (I) and Resistant (R).

a3). It visualizes cross classification results (case study) stored in a database, after compilation of selection results, application of a mathematical model or simultaneous selection index and classification algorithm, where the genetic value or classification of the cross is obtained in eight categories (13). 0-no evaluation; 1-moderately discarded; 2-moderately discarded; 3-discarded; 4-exploratory; 5-moderately proven; 6-proven and 7-very proven. In this case, the C86-12 x CP70-1133 cross was very proven and proven in Villa Clara and Santiago de Cuba provinces, respectively, and without evaluation in Holguín province.

a4). It reflects the selection summaries of the seven databases collected in each series, stages and territory in which it participated. It was possible to determine that 90 individuals were planted in the study crossing, ten did not survive, 80 were evaluated, with 14 selected for a 17.5 % of selection. The rest of individuals were eliminated due to charcoal (7), brown rust (1), low brix (1), flowering (9), vigor (43) and other causes (4).

b1). In the second part of the table it was possible to determine if the cross belongs to the national crossing program. In this case, the C86-12 x CP70-1133 cross is included in the crossing program (CP).

b2) it shows the summaries of the basic statistics and the genetic-statistical parameters carried out on the

Table 1. Background of selection results of the C86-12 x CP70-1133 cross and participating databases (big data approach)

A) Colección de datos por series, etapa y provincia (a1), pruebas de resistencia (a2), clasificación de cruces (a3), resumen de selección (a4)

(a1)						(a2)					(a3)	(a4)									
						Reacción a:					Categ.	Individuos				% de individuos afectados por:					
No.	Año	Etapa	Territ.	Cruce	Combinación	Roya	Carb.	Escal.	VMCA		SASEL	Mtos	Eval.	Selec.	%	Carbon	Roya	Bris	Flores	Vigor	Otras
1	2008	LC1	HL	380	C86-12 x CP70-1133	S x S	I x S	I x S	I x R		S6 H0 V7	6	10	2	20.0				10.0	60.0	10.0
2	2008	LC2	HL	380									2			100.0					
3	2008	LC1	OS	378									14	5	35.7	28.6		14.3	7.1	7.1	7.1
4	2008	LC2	OS	378								1	1				100.0				
5	2009	LC1	HL	540								2	30	2	6.7					86.7	6.7
6	2009	LC2	HL	540								1	1	1	100.0						
7	2011	LC1	VC	450									22	4	18.2	4.5			31.8	45.5	
	Total: 7											10	80	14	17.5	8	1	3	11	54	8

B) Programa nacional de cruces (b1), parámetros genéticos-estadísticos (b2)

(a1)						(b1)	(b2)													
							Individuos			Promedio				Diferencial de selección				Observaciones		
No.	Año	Etapa	Territ.	Cruce	Combinación	PC	Eval.	Selec.	%	Bris	Diám.	Long.	Tallos	Bris	Diám.	Long.	Tallos	Bris	Media	Ob./Sel.
1	2008	LC1	HL	380	C86-12 x CP70-1133	PC	10	2	20.0	23.40	3.00	1.94	12	0.28	0.03	0.14	2	2	2	1
2	2008	LC2	HL	380			2			23.00	2.60	2.48	41	-0.19	-0.01	0.20	22	2	2	2
3	2008	LC1	OS	378			14	5	35.7	23.42	2.52	2.55	12	-0.30	0.07	0.25	-1	5	5	1
4	2008	LC2	OS	378			1													
5	2009	LC1	HL	540			30	2	6.7	23.00	2.90	1.68	11	0.75	0.27	-0.18	5	2	2	1
6	2009	LC2	HL	540			1	1	100.0	23.30	2.60	2.25	28	-0.07	-0.02	0.12	0	1	1	1
7	2011	LC1	VC	450			22	4	18.2	24.21	2.66	2.31	10	0.38	0.36	0.30	-2	4	7	1
Total: 7					Individuos:		80	14	17.5									16	19	

Territ. - Territorios; OS- Oriente Sur; HL- Holguin ; VC- Villa Clara; Eval.- Evaluados; Selec. - Seleccionados;Diám.- Diámetro (cm); Long.- Longitud (m)
PC - Programa nacional de cruces; F- Femenino; M- Masculino; Ob./Sel. - Relación entre el total de observaciones y los seleccionados para el bris
Categ. SASEL - Provincia (S- Santiago de Cuba, H- Holguin y V- Villa clara) y clasificación de cruces desde muy descartado (1) a muy comprobado (7)
Observ. - Cantidad de evaluaciones de bris y media de observ. de diám., long. y tallos; Mtos - Muertos

seven databases collected in each series, stages and territories in which it participated. The mean refractometric brix, diameter and length of stems, as well as the number of stems per seedling were determined. On the other hand, it visualizes the selection differential of the cross under study with respect to the control for each variable evaluated, which shows the progress or response to selection. In this case it was a cross with good results due to most of positive values of the differential selection or very close to zero reached in all the variables evaluated. The number of observations from which the mean values of the variables are derived is also shown as a reference of the sample size.

The big data management model designed to process the volume, variability, speed and veracity of the data generated in the sugarcane genetic selection program shows its potential in the case study of the C86-12 x CP70-1133 cross. In a relatively short period, all the information available on the subject is accessed and makes it possible to recommend its best management.

Many developed countries have begun the application of massive data analysis in precision agriculture and integrated it into traditional production methods (13). These applications have focused on data collection from sensors on high-tech machines, automatic weather stations and satellite information to improve efficiency in yield estimation.

The Australian sugar industry has recognized the potential of the big data approach and has invested heavily in this area of research. These technologies have been incorporated into key areas such as precision agriculture, plant breeding, and geospatial data centers for research and extension (14).

In Australian sugar mills, a *big data* model has been used to relate weather variables to crop productivity, as well as for yield prediction (14,15). Another study developed in China allowed monitoring and estimating yield using the internet of things, sensor technologies and image processing (16). The results showed that crop growth could be visualized, along with temperature and humidity variables, so that growers can implement remote visual management and improve their production efficiency.

This research and the results of the present work support the use of data mining and *big data* technologies to facilitate decision-making and contribute to the improvement of the sugar agribusiness. Most importantly, the approach described in this paper can be easily extended to other areas of crop research and agricultural industries to recommend better farming practices.

CONCLUSIONS

- A management model was obtained with a macro data approach, assisted by the SASEL interface, to facilitate decision-making in the Sugarcane Breeding Program in Cuba and increase its efficiency.
- Through this information management model, it was possible to obtain a compilation of primary data, selection results, estimate the genetic value of parents and crosses and classify the crosses used in the Breeding Program.

BIBLIOGRAPHY

1. Sun J, Zhou Z, Bu Y, Zhuo J, Chen Y, Li D. Research and development for potted flowers automated grading system based on internet of things. Journal of Shenyang Agricultural University [Internet]. 2013;44(5):687-91. Available from: https://www.cabdirect.org/cabdirect/abstract/20133406244
2. Yang C. Big Data and its potential applications on agricultural production. Crop, Environment &

Bioinformatics [Internet]. 2014;11(1):51-6. Available from: https://www.cabdirect.org/cabdirect/abstract/20143190953

3. Chen M, Mao S, Liu Y. Big data: A survey. Mobile networks and applications [Internet]. 2014;19(2):171-209. Available from: https://dl.acm.org/doi/10.1007/s11036-013-0489-0
4. Mazzocchi F, Lapointe FJ. Sobre el 'big data':¿ Cómo podríamos dar sentido a los macrodatos? Mètode: Revista de difusión de la Investigación [Internet]. 2020;1(104):34-41. Available from: https://dialnet.unirioja.es/servlet/articulo?codigo=7391781
5. Wolfert S, Ge L, Verdouw C, Bogaardt M-J. Big data in smart farming-a review. Agricultural systems [Internet]. 2017;153:69-80. Available from: https://www.sciencedirect.com/science/article/pii/S0308521X16303754
6. Park S, Jackson P, Berding N, Inman-Bamber G. Conventional breeding practices within the Australian sugarcane breeding program. In: Proceedings of the Australian Society of Sugar Cane Technologists [Internet]. 2007. p. 113-21. Available from: https://www.researchgate.net/profile/Nils-Berding/publication/305390103_Conventional_breeding_practices_within_the_Australian_sugarcane_breeding_program/links/578c603608ae5c86c9a14e99/Conventional-breeding-practices-within-the-Australian-sugarcane-breeding-program.pdf
7. Yadav S, Jackson P, Wei X, Ross EM, Aitken K, Deomano E, et al. Accelerating genetic gain in sugarcane breeding using genomic selection. Agronomy [Internet]. 2020;10(4):585. Available from: https://www.mdpi.com/2073-4395/10/4/585
8. González R. Variedades de caña de azúcar cultivadas en Cuba. Cronología, legislación, metodologías y conceptos relacionados. Instituto Cubano de Investigaciones de Derivados de La Caña de Azúcar; 2019.
9. Rodríguez R, Puchades Y, Abiche W, Rill S, García H. SASEL: software for data management generated in the Cuban sugarcane-breeding program. In: Proceedings of the International Society of Sugar Cane Technologists. 2016. p. 63-6.
10. Jorge H, González R, Casas M, Jorge I. Normas y procedimientos del programa de mejoramiento genético de la caña de azúcar en Cuba. PUBLINICA, La Habana. 2011;
11. Rodríguez-Gross R, Puchades-Isaguirre Y, Aiche-Maceo W, Cornide-Hernández MT. Modelo matemático para estimar el valor genético de progenitores y cruces en caña de azúcar. Cultivos Tropicales [Internet]. 2018;39(2):81-8. Available from: http://scielo.sld.cu/scielo.php?pid=S0258-59362018000200011&script=sci_arttext&tlng=en
12. Rodríguez-Gross R, Puchades-Isaguirre Y, Aiche-Maceo W. Metodología de validación y manejo de cruces en la mejora genética en caña de azúcar. Cultivos Tropicales [Internet]. 2020;41(1). Available from: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0258-59362020000100002
13. Kumar H, Menakadevi T. A review on big data analytics in the field of agriculture. International Journal of Latest Transactions in Engineering and Science [Internet]. 2017;1(4):1-10. Available from: http://www.ijltes.com/wp-content/uploads/2017/02/1.pdf
14. Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for sustainable development [Internet]. 2016;36(2):27. Available from: https://link.springer.com/content/pdf/10.1007/s13593-016-0364-z.pdf
15. Everingham Y, Sexton J, Robson A. A statistical approach for identifying important climatic influences on sugarcane yields. In: Proceedings of the 37th Conference of the Australian Society of Sugar Cane Technologists, 28-30 April 2015, Bundaberg, Queensland, Australia [Internet]. Australian Society of Sugar Cane Technologists; 2015. Available from: https://www.researchgate.net/publication/287208302_A_Statisical_Approach_for_identifying_Important_Climatic_Influences_on_Sugarcane_Yields
16. Biqing L, Yongfa L, Miao T, Shiyong Z. Design and Implementation of Sugarcane Growth Monitoring System based on RFID and ZigBee. International Journal of Online Engineering [Internet]. 2018;14(3). Available from: https://www.researchgate.net/publication/324114641_Design_and_Implementation_of_Sugarcane_Growth_Monitoring_System_based_on_RFID_and_ZigBee