ISSN impreso: 0258-5936
ISSN digital: 1819-4087

Ministerio de Educación Superior. Cuba
Instituto Nacional de Ciencias Agrícolas
http://ediciones.inca.edu.cu

INFORMATION SYSTEM FOR THE CONTROL OF THE VARIETY BIODIVERSITY AND CHARACTERIZATION OF BEAN CULTIVATION IN CUBA

Sistema de información para el control de la Biodiversidad de variedades y caracterización del cultivo de frijol en Cuba

Nelson Verdesia Hernández✉, Ariel Hernández Musa, Irina Blanco Gil and Alexis Lamz Piedra

ABSTRACT. In the Local Agricultural Innovation Project (PIAL according its acronym in Spanish), the study of the genetic diversity of germplasm banks of agricultural crops is of great importance for the use and conservation of plant genetic resources in Cuba. The provision of farmers in Cuba of the widest diversity of species and varieties of crops has been without doubt one of its basic purposes. The PIAL does not have an automated system for the control of the processes linked to the thematic axis of genetic diversity and technology. That is to say, the information of the composition of the diversity is stored manually, as well as the process of assessing the behavior agro-morphology from the characterization of the variability in lines of common bean (*Phaseolus vulgaris* L) sown in late season. Due to the foregoing raised proposes to carry out a system that will automate the control of information in the agricultural biodiversity. The design of the system is based on the RUP methodology and their implementation was carried out using tools of the Java platform. The use of the system facilitates the control agricultural biodeversity information. In addition of saving resources and streamline the work of the researchers in the search for bean varieties that have better performance out of the whole year.

Key words: crops, disemination, germplasm, seeds, free software

RESUMEN. En el Proyecto de Innovación Agropecuaria Local (PIAL) el estudio de la diversidad genética de los bancos de germoplasma de cultivos agrícolas tiene gran importancia para el uso y conservación de los recursos fitogenéticos en el país. La puesta a disposición de los agricultores en Cuba de la más amplia diversidad de especies y variedades de cultivos ha constituido sin dudas uno de sus propósitos básicos. El PIAL no cuenta con un sistema automatizado para el control de los procesos vinculados al eje temático de Diversidad Genética y Tecnológica, la información de la composición de la diversidad se almacena de forma manual, así como el proceso de la evaluación del comportamiento agro-morfológico, a partir de la caracterización de la variabilidad en líneas de frijol común (*Phaseolus vulgaris* L) sembradas en época tardía. Debido a lo anterior planteado se propone realizar un sistema que automatice el control de información de la biodiversidad agrícola. El diseño del sistema está basado en la metodología RUP y su implementación se realizó mediante herramientas de la plataforma Java. La utilización del sistema facilita el control de información de la biodiversidad agrícola, además de ahorrar recursos y agilizar el trabajo de los investigadores en la búsqueda de variedades de frijol que tengan un mayor rendimiento en todo el año.

Palabras clave: cultivos, diseminación, germoplasma, semillas, software libre

INTRODUCTION

The study of the genetic diversity of germplasm banks of agricultural crops is of great importance for the use and conservation of plant genetic resources.

Knowing and understanding the diversity structure of local varieties is vital in the identification of those populations that should be conserved in the best places for the collection of germplasm, and for tracking changes in diversity patterns over the course of *in situ* conservation practices. This requires the characterization of the diversity in gene pools and gene banks, to extend and complement the characterization based on morphological descriptors

Instituto Nacional de Ciencias Agrícolas (INCA), Carretera Tapaste, Km 3½, Gaveta Postal 1. San José de las Lajas, Mayabeque. Cuba. CP 32700
✉ nelson@inca.edu.cu

as well as biochemical and molecular markers. In addition, knowing the available diversity allows a more efficient use of these phytogenetic resources, since it recognizes ample and diverse sources of improvement, materials and populations that carry genes for characters of interest in the improvement of plants, which makes it possible to incorporate genetic diversity and achieve a greater gain in the characters of agronomic value (1).

The provision of the widest diversity of species and varieties of crops to farmers in Cuba has undoubtedly been one of the basic purposes of PIAL, as well as the diversification of agricultural systems focused on FP (Participatory Plant Breeding) in its beginnings and later the PIAL. This practice showed in Cuba evidences of the importance of the active participation of the farmers in the decision making about the selection, multiplication, maintenance and conservation of the phytogenetic resources at local level, in very close collaboration with researchers of the National Institute of Agricultural Sciences (INCA) and other scientific research institutions and the education sector of the country (2).

The large volume of information used to execute the aforementioned processes is done manually, causing a decrease in the speed and efficiency in the handling of the information, since the existing CDBAs (Agricultural Biodiversity Dissemination Center) are located in 28 different municipalities and 10 provinces of the country, which hinders access to information.

Currently PIAL does not have an automated system for the control of processes linked to the thematic axis of Genetic and Technological Diversity; the composition of the diversity is stored manually as the operation of the CDBA.

Due to the above, it is proposed to develop a system that automates the operation of the CDBAs in Cuba with the objective of controlling the information on the biodiversity of the crop varieties and characterizing the varieties of the bean crop.

In order to solve the problem described, the objective is to develop the "SCIDBA" system to promote an adequate control of the biodiversity information of varieties and the characterization of the bean crop.

MATERIALS AND METHODS

The tools, technologies and methodologies used for the development of the module were selected by the project's architecture group, taking into account that they are the most appropriate for the work to be done following the country's policy of migrating to free software.

Groovy as a language for web development

Groovy is an object-oriented programming language implemented on the Java platform. It has characteristics similar to Python, Ruby, Perl and Smalltalk. The JSR 241 specification is responsible for standardizing it for future inclusion as an official component of the Java platform (3).

Grails framework

Grails is a dynamic framework for the development of web applications on the Java platform, which follows the principles ***Do not repeat yourself*** and ***Convention over configuration***. Grails is more than just a Model View Controller (MVC) framework, it also offers persistence layer, service layer, servlet container and database manager. It is based on several well-known and tested Java frameworks and libraries such as *Spring Framework, Hibernate, Sitemesh, Log4j, Jetty, Hsqldb* (4).

Integrated Development Environment (IntelliJ IDEA CE)

IntelliJ IDEA is one of the best IDEs (Integrated Development Environment) for Java, by the hand of JetBrains that always had features that we do not find in others (Eclipse, NetBeans, among others) such as automatic saving and compilation, control of suggestions and highlighting of code within even strings, the most intelligent self-completed code (5).

Visual designer of dynamic reports

iReport is a visual open source designer for JasperReports written in Java. It is a program that helps users and developers who use the JasperReports library to design reports visually. Through a rich and simple interface to use iReport, it provides the most important functions to create reports in a short time. iReport can help people who do not know the XML syntax to generate JasperReports reports (6).

PostgreSQL database management systems

It is an object-relational database management system, distributed under BSD license and with its source code freely available. It uses a client/server model and it uses multi-processes instead of multithreading to guarantee the stability of the system. A failure in one of the processes will not affect the rest and the system will continue to work. The last series of production is 9.3.

Its technical characteristics make it one of the most powerful and robust databases in the market. Its development began more than 16 years ago, and

during this time, stability, power, robustness, ease of administration and implementation of standards have been the characteristics that have been most taken into account during its development. PostgreSQL works very well with large amounts of data and a high concurrency of users accessing the system at the same time (7).

Environment for knowledge analysis of the University of Waikato

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a data set or called from their own Java code. *Weka* contains tools for data pre-processing, classification, regression, grouping algorithm, association rules, and visualization. It is also very suitable for the development of new machine learning schemes, it is software published under the GNU General Public License (8).

JavaScript library for interactive maps

Leaflet is open source a modern JavaScript library for friendly mobile interactive maps. It is developed by Vladimir Agafonkin with a team of dedicated collaborators.

The library is designed with simplicity, performance and ease of use in mind. It works efficiently through all the main desktop and mobile platforms of the box, taking advantage of HTML5 and CSS3 in modern browsers while still being accessible (9).

JavaScript library

JQuery is a JavaScript library that allows you to simplify the way you interact with HTML documents, manipulate the DOM tree, manage events, develop animations and add interaction with the AJAX technique to web pages, much simpler with an easy to use API that works through a multitude of browsers. With a combination of versatility and expandability, jQuery has changed the way millions of people write JavaScript (10).

Description of the algorithm to be used

The main components analysis is to build a set of new variables or components, with the characteristic that in this set most of the information or initial variability will be concentrated in the first axes or components. This result allows in turn reducing the dimensionality of the problem, facilitating the characterization of the elements of the sample and the search for correlation structures between variables.

These new variables or components are no more than linear combinations of the original variables. They are constructed in such a way that there is no correlation between them; In addition, they have the characteristic that each one has maximum variance, that is, it explains as much initial information as possible.

Notions of the mathematical foundation

As in any multivariate method, we start from the initial data matrix X:

$$X=\begin{bmatrix} & \vdots & \\ \dots & ij & \dots \\ & \vdots & \end{bmatrix}_{nxp}$$

Thus, the element ij of the matrix represents the observed value of the variable j in the individual i. In this case, it is appropriate to point out that the p variables must be of a continuous nature, since the method works with the Pearson correlation coefficient, designed to measure the existing linear relationship between continuous variables.

From this matrix, we estimate the vector of means upx1 = (u1, u2, ..., up) and the matrix of variances and co-variances Epxp by means of X means px1 and Spxp respectively.

The objective is to find p functions (Y1, Y2, ..., Yp), which are expressed as a linear combination of the original variables, which are called principal components.

Being:

$$Y_1 = \sum_{j=1}^{p} A_{1j} X_j, \ldots, Y_p = \sum_{j=1}^{p} A_{pj} X_j$$

Thus, to find Y1, that is, the first component, it is necessary to find the coefficients A1j j = 1..p, so that the variance of Y1 is maximum, subject to the condition:

$$\sum_{j=1}^{p} A_{1j} = 1$$

Which ensures the uniqueness of the solution.

To find Y2 (second component), it is necessary to find the coefficients A2j j = 1..p, so that the covariance of Y2 with Y1 is equal to zero; In addition, it must be fulfilled that the variance of Y2 is maximum and that:

$$\sum_{j=1}^{p} A_{2j} = 1$$

Note that in the case of the calculation of Y2, one more condition is required; it is for this reason that it must be obtained that the variance of Y2 is going to be less than or equal to the variance of Y1.

To find Y3 (third component), it is necessary to find the coefficients A3j j = 1..p, so that the covariance of Y3 with Y2 and the covariance of Y3 with Y1 are both equal to zero. In addition, Y3 must have maximum variance and the coefficients must meet the condition:

$$\sum_{j=1}^{p} A_{j} = 1$$

For the same reason, the variance of Y3 must be less than or equal to the variance of Y2 and Y1. The rest of the components are calculated by the same algorithm, until arriving at the component p.

The solution to this problem is to find the values and eigenvectors of the S matrix of variances and covariances. Thus, for example, the largest of the eigenvalues of S will be the value corresponding to the variance of Y1, and its associated eigenvector is identified with the coefficients A1j, j = 1..p. The second eigenvalue (establishing a decreasing order) will be the value corresponding to the variance of Y2 and its associated eigenvector is represented by the coefficients A2j, j = 1..p; and so on until you get to Yp. Some authors offer in detail the demonstration of this result (11,12).

Note that finally the components fulfill the property of being incorrect and the variance of the first will be greater than that of the second and so on.

The sum of the variances of all the components will be equal to the trace of the S matrix of variances and covariances, because these variances are not more than the eigenvalues. Now, this is not more than the sum of the variances of the original variables Xi, i = 1..p, since these are the elements of the diagonal S.

Var(Y1)+Var(Y2)+...+
Var(Yp)=Traza(S)=Var(X1)+Var(X2)+...+Var(Xp)

It is for this result that the Analysis of Principal Components can be used in the reduction of dimensionality, that is, tries to explain with fewer components the initial information that is collected in the initial matrix of data X.

Once the components are constructed, the correlation of each component with the initial variables is calculated; this is a very important step, since from these correlations, it is that you will have a criterion to characterize the axes or components.

It is appropriate to point out that some authors suggest that, on occasion, it is more effective instead of calculating values and eigenvectors from the S matrix of variances and covariances, calculating them from the correlation matrix. It is suggested that the latter should be used when the original variables are measured on a different scale, since the data are standardized through the calculation of the correlation matrix (13,14).

RESULTS AND DISCUSSION

The system is directly linked to the thematic axis of Genetic and Technological Diversity of the PIAL in order to improve and facilitate the control of information on the biodiversity of varieties and the characterization of the bean crop in Cuba for this is based on the use of analysis statistics such as the mean, standard deviation, coefficient of variation and the performance of principal components analysis. The process of the evaluation of the agro-morphological behavior from the characterization of the variability in lines of common bean (*Phaseolus vulgaris* L) planted in late season begins with the planting of all varieties of the same in a “Random Blocks” design. “The evaluation is done using 14 agro-morphological characters that include phenological, morphological parameters, performance and its components, and resistance to rust (*Uromyces appendiculatus*). To evaluate the genotypes, 14 morphoagronomic variables were taken into account, according to the descriptors recommended for the characterization of bean genotypes (Table 1) (15).

To carry out the process of evaluation of the agro-morphological behavior of the bean crop, the Characterization option should be selected where the variety to be processed is chosen. Once selected, the list of varieties is shown, as well as the option to add a new variety, modify, eliminate and the analysis option (Figure 1).

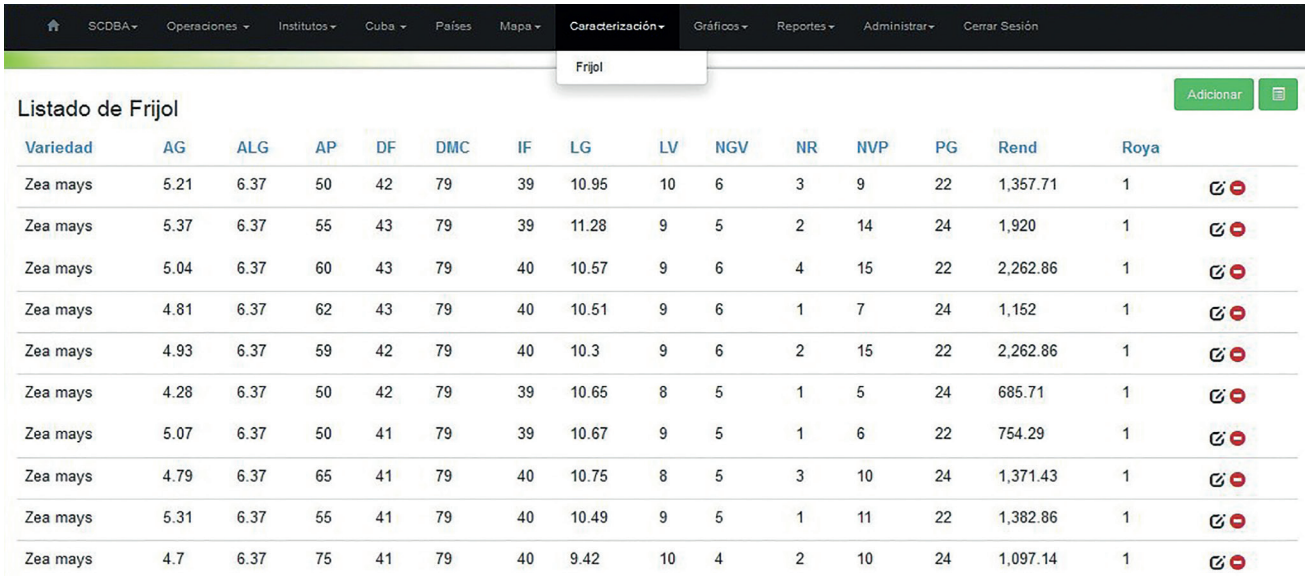
Table 1. Variables evaluated for the evaluation of the agro-morphological behavior of common bean lines

No.	Code	Variables	No.	Code	Variables
1	AP	Height of the plant (cm)	8	NGV	Number of grains per pod
2	NR	Number of branches	9	PG	Weight of 100 grains
3	LV	Pod length (cm)	10	Rend.	Yield (t ha^{-1})
4	IF	Days at the beginning of flowering	11	LG	Pod length
5	DF	Days to flowering	12	AG	Grain width
6	DMC	Days to harvest maturity	13	Al.G	Grain height
7	NVP	Number of grains per pod	14	R	Incidence of rust (*Uromyces apendiculatus*)*

*Scale that classifies the reaction of germplasm to the pathogen of rust in three discrete categories: resistant, intermediate or susceptible (14)

Once the above data is inserted, the results obtained after applying the Principal Component Analysis algorithm (Figure 2) are shown, where the data that are independent of the root and its value is greater than or equal to 0.5 are those that they have a greater correlation between them.

Figure 3 shows the main components in which the values that are taken are greater than 1 and the rest are discarded. With these results the researcher or the specialist can appreciate which of the varieties of bean crop is the correct one for planting in later times.



Variedad	AG	ALG	AP	DF	DMC	IF	LG	LV	NGV	NR	NVP	PG	Rend	Roya
Zea mays	5.21	6.37	50	42	79	39	10.95	10	6	3	9	22	1,357.71	1
Zea mays	5.37	6.37	55	43	79	39	11.28	9	5	2	14	24	1,920	1
Zea mays	5.04	6.37	60	43	79	40	10.57	9	6	4	15	22	2,262.86	1
Zea mays	4.81	6.37	62	43	79	40	10.51	9	6	1	7	24	1,152	1
Zea mays	4.93	6.37	59	42	79	40	10.3	9	6	2	15	22	2,262.86	1
Zea mays	4.28	6.37	50	42	79	39	10.65	8	5	1	5	24	685.71	1
Zea mays	5.07	6.37	50	41	79	39	10.67	9	5	1	6	22	754.29	1
Zea mays	4.79	6.37	65	41	79	40	10.75	8	5	3	10	24	1,371.43	1
Zea mays	5.31	6.37	55	41	79	40	10.49	9	5	1	11	22	1,382.86	1
Zea mays	4.7	6.37	75	41	79	40	9.42	10	4	2	10	24	1,097.14	1

Figure 1. Characterization of the bean crop

Matriz de Correlaciones

	AG	AP	NR	IF	DF	DMC	NVP	NGV	LV	PG	REND	LG	ALG
AG	1.0	0.08	-0.03	-0.14	0.06	0.15	0.12	-0.04	0.24	0.66	0.32	0.49	0.5
AP	0.08	1.0	0.17	-0.43	-0.41	-0.1	-0.07	0.07	0.68	0.08	0.07	0.34	0.15
NR	-0.03	0.17	1.0	0.12	0.1	0.15	0.37	0.08	0.12	0.02	0.25	-0.05	-0.2
IF	-0.14	-0.43	0.12	1.0	0.84	0.27	0.04	0.02	-0.49	-0.24	-0.14	-0.41	-0.34
DF	0.06	-0.41	0.1	0.84	1.0	0.63	0.25	0.09	-0.43	-0.02	0.16	-0.41	-0.17
DMC	0.15	-0.1	0.15	0.27	0.63	1.0	0.42	0.08	-0.19	0.11	0.41	-0.32	-0.02
NVP	0.12	-0.07	0.37	0.04	0.25	0.42	1.0	0.08	0.06	0.32	0.73	-0.03	0.03
NGV	-0.04	0.07	0.08	0.02	0.09	0.08	0.08	1.0	0.09	-0.22	0.23	-0.27	-0.11
LV	0.24	0.68	0.12	-0.49	-0.43	-0.19	0.06	0.09	1.0	0.35	0.26	0.43	0.11
PG	0.66	0.08	0.02	-0.24	-0.02	0.11	0.32	-0.22	0.35	1.0	0.62	0.57	0.33
REND	0.32	0.07	0.25	-0.14	0.16	0.41	0.73	0.23	0.26	0.62	1.0	0.15	0.11
LG	0.49	0.34	-0.05	-0.41	-0.41	-0.32	-0.03	-0.27	0.43	0.57	0.15	1.0	0.53
ALG	0.5	0.15	-0.2	-0.34	-0.17	-0.02	0.03	-0.11	0.11	0.33	0.11	0.53	1.0

Figure 2. Correlation Matrix

Componentes

Valor específico	Proporción	Acumulado
3.68582	0.28352	0.28352
2.85963	0.21997	0.5035
1.76886	0.13607	0.63956
1.03466	0.07959	0.71915
0.9202	0.07078	0.78994
0.7311	0.05624	0.84617
0.69471	0.05344	0.89961
0.4556	0.03505	0.93466
0.28105	0.02162	0.95628

Figure 3. Main Components

CONCLUSIONS

- The different existing computer systems do not meet current needs, so it was demonstrated the need to develop a control system for biodiversity of varieties and characterization of bean cultivation in Cuba to support decision making.
- The Control System of the biodiversity information of varieties and characterization of bean cultivation in Cuba was analyzed, designed and implemented, which allows to efficiently managing all the processes that are carried out in the Agricultural Diversity group, favoring to the decision making of researchers and specialists.
- The algorithm selected is the correct one because it has advantages such as speed, which can be considerable when dealing with large volumes of data.
- Through the use of the control system of biodiversity information on varieties and characterization of bean cultivation in Cuba, it was shown that the PIAL had a significant economic savings as well as a reduction in processing times and increased quality of work demonstrating that the system is feasible.

BIBLIOGRAPHY

1. Ortiz PHR, Miranda LS, Martínez CM, Ríos LH, Cárdena TRM, de la Fe MCF, *et al.* La Biodiversidad Agrícola en manos del campesinado cubano. La Habana, Cuba: Ediciones INCA; 2013. 357 p.
2. Ortiz PR, Angarica L, Acosta RR, Guevara HF. Manual de Monitoreo y Evaluación Participativos con enfoque de Género. La Habana, Cuba: Ediciones INCA; 2014. 126 p. (Programa de Innovación Agropecuaria Local, PIAL III).
3. Dickinson J. Grails 1.1 web application development. Packt Publishing Ltd; 2009. 310 p.
4. Layka V, Judd CM, Nusairat JF, Shingler J. Deploying and upgrading grails applications. In: Beginning Groovy, Grails and Griffon. Berkeley, CA: Springer Link; 2013 [cited 2018 Apr 9]. p. 291–303. doi:10.1007/978-1-4302-4807-1_12
5. Contreras C. Manual de Ireport [Internet]. Scribd. [cited 2018 Apr 9]. Available from: https://es.scribd.com/doc/37388195/Manual-de-Ireport
6. Martínez R. Sobre PostgreSQL [Internet]. PostgreSQL-es, Portal en español sobre PostgreSQL,[En línea] Octubre. 2010 [cited 2018 Apr 9]. Available from: www. postgresql-es. org. Sobre PostgreSQL.
7. Weka 3 - Data Mining with Open Source Machine Learning Software in Java [Internet]. WEKA The University of Waikato. [cited 2018 Apr 9]. Available from: https://www.cs.waikato.ac.nz/ml/weka/
8. Agafonkin V. Leaflet-a JavaScript library for interactive maps [Internet]. Online: http://leafletjs. com. 2016. Available from: http://mourner.github.io/Leaflet/download.html
9. Cooley WW, Lohnes PR. Multivariate data analysis. Biometrische Zeitschrift. 1971;15(4):364. doi:10.1002/bimj.19730150413
10. Bibeault B, Katz Y, De Rosa A. jQuery in Action, Third Edition. Shelter Island, NY: Manning Publications; 2015. 504 p.
11. Gnanadesikan R. Methods for statistical data analysis of multivariate observations. New York: Wiley; 1977. 311 p.
12. Morrison D. Multivariate Statistics [Internet]. New York: John Wiley and Sons; 1979 [cited 2018 Apr 9]. 414 p. Available from: http://www.multivariatestatistics.org/
13. Varela M. Análisis multivariado de datos. Aplicación a las Ciencias Agrícolas. 1998.
14. Schoonhoven AV, Corrales P. Sistema estándar para la evaluación de germoplasma de frijol. [Internet]. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia; 1987. 56 p. Available from: https://cgspace.cgiar.org/handle/10568/69699
15. Lamz Piedra A, Cárdenas Travieso RM, Ortiz Pérez R, Montero Tavera V, Martínez Coca B, de la Fé Montenegro CF, *et al.* Evaluación del comportamiento agro-morfológico a partir de la caracterización de la variabilidad en líneas de frijol común (*Phaseolus vulgaris* L.) sembradas en época tardía. Cultivos Tropicales. 2016;37(2):108–14.

Received: June 17th, 2017
Accepted: February 19th, 2018

Revista Cultivos Tropicales
Instituto Nacional de Ciencias Agrícolas