

SELDESCRIP V1.0: APLICACIÓN WEB PARA LA SELECCIÓN DE DESCRIPTORES DE BANCOS DE GERMOPLASMA VEGETAL

Osmany Molina Concepción*, Adrián Peña Ángel, Yaselis Guillén López y Carmen C. Pons Pérez.

Instituto de Investigaciones de Viandas Tropicales (INIVIT), Cuba. INIVIT, Apartado 6, Santo Domingo, CP: 53 000, Villa Clara, Cuba.

* Autor para correspondencia: taxonumeric@inivit.cu, <https://orcid.org/0000-0003-4199-6652>.

Recibido: 29 de abril de 2022; Aceptado: 15 de junio de 2022

RESUMEN

La obtención de un amplio número de variables descriptoras en bancos de germoplasma vegetal puede, *a priori* parecer deseable, pero probablemente no todos los descriptores sean relevantes para las tareas de clasificación, y teniendo en cuenta que la mayoría de los mejoradores no tienen un manejo fluido de los paquetes estadísticos de R, se creó una Interfaz gráfica en R con *Shiny* que facilite realizar la selección de descriptores. Para cumplir este objetivo se recurre al uso de herramientas de inteligencia artificial y estadísticas implementadas en el *Software R* para la selección de variables. Esta interfaz tendrá potencialidad e independencia al ser este un lenguaje con entorno de programación libre. Además, permite a los usuarios interactuar con sus datos sin requerir conocimientos en programación.

Palabras clave: Aplicación web, *Shiny*, *Software R*.

SELDESCRIP V1.0: WEB APPLICATION FOR THE SELECTION OF DESCRIPTORS IN VEGETABLE GENBANKS

ABSTRACT

The obtention of a wide number of descriptive variables in vegetal germplasm banks could at first seem desirable, but probably not all the descriptors are relevant for classification tasks, and taking into account that most of the improvers do not have a fluent handling of the statistical packages of R, a graphical interface was created in R with *Shiny*, which facilitates the selection of descriptors. To meet this objective, artificial intelligence and statistical tools are implemented in R software for the selection of variables. This interface will have potential and independence as this is a free programming environment language. In addition, it allows the users to interact with their data without requiring programming knowledge.

Keywords: Web Application, *Shiny*, *Software R*.

INTRODUCCIÓN

Mediante la caracterización y evaluación de los bancos de germoplasma vegetal, los genetistas proveen información sobre la diversidad genética contenida en estos. La obtención de un amplio número de variables descriptoras puede *a priori* parecer deseable, pero probablemente no todos los descriptores sean relevantes para las tareas de clasificación; por lo cual, una de las dificultades que se deben enfrentar en el momento de una clasificación es la dimensión de los datos. Una forma de resolver el problema es considerar técnicas de selección de variables. Para cumplir este objetivo

se recurre al uso de métodos de selección de variables y de herramientas de inteligencia artificial y estadísticas en el *Software R* (R Core Team, 2021). R es un lenguaje de programación de código abierto utilizado por muchos en las ciencias agrícolas y biológicas (Elliott y Elliott, 2020; Lihua et al., 2022).

A diferencia de la mayoría de los programas que tienen interfaces tipo ventana, R es manejado a través de una consola en la que se introduce comandos propios de su lenguaje para acceder a todos los procedimientos; sin embargo, es posible diseñar interfaces gráficas para facilitar la interacción con los usuarios no familiarizados con su consola (Salas, 2008).

Mediante el paquete *R-Commander*, los usuarios pueden realizar el procesamiento de datos estadísticos fácilmente a través de los menús proporcionados en la ventana de *R-Commander*, tan fácil como usar SPSS (Landau y Everitt, 2004), Minitab (Moler y Mathworks, 2012) u otro *software* estadístico con licencia. Con *R-Shiny* se han creado muchas aplicaciones basadas en la *web* enfocadas a numerosos ámbitos de la investigación (Beeley, 2013; Arnholt, 2018). Debido a las ventajas de este, se puede acceder a él a través de la *web* sin tener que utilizar un ordenador que contenga R, ni dominar el *software* estadístico (Satyahadewi y Perdana, 2021).

R es una herramienta muy flexible y potente para analizar y visualizar datos. *Shiny* es el compañero perfecto de R, ya que permite compartir de forma rápida y sencilla, análisis y gráficos de R con los que los usuarios pueden interactuar y consultar a través de la *web* (Beeley y Sukhdev, 2018). *Shiny* es un paquete de R que permite crear fácilmente aplicaciones *web* ricas e interactivas (Chang et al., 2020). Algunos paquetes basados en *Shiny* y que puede ejecutarse localmente o en un servidor es *Radiant* (Nijs, 2020), *BotanizeR* (Weigelt et al., 2021), *PhenoExam* (Cisterna et al., 2021), *shinyGISPA* (Dwivedi y Kowalski, 2018), *BioInstaller* (Li et al., 2018) entre otros.

Shiny es la librería de R por excelencia para crear aplicaciones *Web* donde se utiliza la arquitectura Cliente y Servidor, esta librería genera el HTML5/JavaScript/CSS necesario para construir las aplicaciones *web* con parámetros y variables dinámicas que permiten interactuar con los datos y ver diferentes resultados sin necesidad de programar una aplicación *web*.

Las aplicaciones *web* desarrolladas con el paquete *Shiny* funcionan de la misma forma que otras aplicaciones *web*, con la ventaja de que R y *Shiny* genera el código necesario para facilitar la creación de una aplicación *web*, sin necesidad de utilizar tecnologías *web*, para ello genera código HTML que muestra los elementos en la página *web* y ofrece un modelo de interacción entre los componentes interactivo y dinámico.

Por lo cual, el objetivo de este trabajo fue implementar una aplicación *web* para la selección de descriptores en bancos de germoplasma vegetal en lenguaje R con *Shiny*, con vista a ayudar a los investigadores en la interpretación y análisis de los resultados sin necesidad de conocer el lenguaje de programación de R.

MATERIALES Y MÉTODOS

El trabajo se desarrolló por el grupo de Bioinformática del Instituto de Investigaciones de Viandas Tropicales (Inivit), partiendo de un diseño logrado sobre el análisis de los requerimientos, que son el conjunto de técnicas y procedimientos utilizados para conocer los elementos necesarios a definir y especificar su función, vincular la interfaz con otros elementos del sistema y establecer los enlaces de diseño que debe cumplir el programa.

El procesamiento incluye un sistema de asistencia al usuario final que le sugiere los pasos a seguir en el análisis de sus datos y ayuda a interpretar los resultados.

En la figura 1 se presenta el diseño del flujo de trabajo del cual se parte para el desarrollo de la aplicación Web. Para la selección de descriptores mínimos. De manera general, se comienza con el tratamiento de los datos (cargar fichero, normalización, definición del tipo de variable), luego se ejecutan los métodos de reducción de dimensionalidad (análisis de correspondencia múltiple y análisis de componentes principales), alternatively es posible realizar una selección de atributos (*RandomForest*, *Chi-Cuadrado*, *OneR*, *HINoV.Mo*, *CFS*, *Information Gain*). A partir de los conjuntos de datos resultantes se realiza el agrupamiento o análisis de clúster con métodos aglomerativo jerárquicos y particionales. Y por último se validan los agrupamientos haciendo una comparación con las clasificaciones originales.

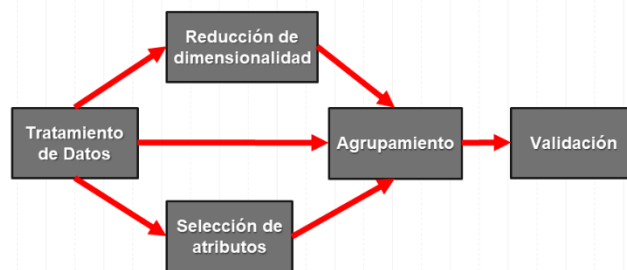


Figura 1. Flujo de la aplicación web.

La programación de los algoritmos y la interfaz se ejecutó en el lenguaje R (R Core Team, 2021) con *Shiny*, para utilizar sus amplias potencialidades al tener disponible en su sitio en Internet una gran cantidad de paquetes con una vasta variedad de funciones que se pueden utilizar en la programación. Además, *Shiny* es una herramienta que facilita crear aplicaciones web interactivas que permiten a los usuarios interactuar con sus datos sin tener que manipular el código.

RESULTADOS Y DISCUSIÓN

Los requerimientos para el diseño fueron expuestos por investigadores que administran los bancos de germoplasma y que tienen cierta experiencia en el desarrollo de los análisis multivariados aplicados a la taxonomía numérica mediante otros programas estadísticos. A partir de esto, fueron seleccionados los análisis que se incluyen separados en: métodos de reducción de datos y clasificación. Dentro de los métodos de reducción de datos se encuentran: análisis de correspondencia múltiple (para variables discretas), análisis por factores (para variables continuas y para variables continuas y discretas), métodos de selección de atributos (para variables continuas y discretas). Posteriormente, el usuario puede seleccionar diferentes medidas de distancia y métodos de conglomeración, lo que proporciona una amplia variedad de clasificaciones diferentes.

En el caso que los datos de entrada sean variables mixtas (continuas y discretas) se implementó como medida de distancia la métrica de *Gower*, y se pueden utilizar cinco métodos de agrupamiento jerárquicos aglomerativos y uno particional. En todos los

casos se proporcionan: el coeficiente aglomerativo, el coeficiente de correlación cofenética, el índice de *Dunn*, el ancho de silueta.

El procesamiento incluye la visualización a través de tablas y gráficos de los resultados. Para lograr este sistema se confrontó con criterios dados por los especialistas de los bancos de germoplasma y se aprovecharon sus conocimientos de estadística aplicada en este campo.

Descripción de la aplicación web

La interfaz implementada contiene una ventana principal con un menú de opciones en la parte superior. Estas opciones son: Fichero, Analizar y Ayuda. Mediante la opción “Fichero” se puede abrir un fichero de datos que se encuentra en algún directorio, el fichero puede tener diferentes formatos tales como: Excel, SPSS, Texto (csv) y Texto (txt), también se muestran un resumen de los datos y de las estructuras de las variables (Figura 2). Conjuntamente en el submenú “Transformaciones” se pueden editar los datos permitiendo al usuario multiplicar, eliminar y transformar variables (Figura 3).

The screenshot displays the initial interface of the web application. At the top, there is a navigation bar with three tabs: 'Fichero', 'Analisis', and 'Conglomerados'. The 'Fichero' tab is active. On the left side, there is a sidebar for file upload settings. It includes a 'Data del Usuario' section with a 'Browse...' button and a text input field containing 'xanthosoma_var_nominal-ld-csv.csv'. Below this is an 'Upload complete' button. The 'Tipo de datos' section has a dropdown menu set to '.csv'. The 'Separator' section has radio buttons for 'Comma' (selected), 'Semicolon', and 'Tab'. The 'Quote' section has radio buttons for 'None', 'Double Quote' (selected), and 'Single Quote'. The 'Display' section has radio buttons for 'Head' (selected) and 'All'. A 'Save' button is at the bottom of the sidebar. The main content area has a 'View' button set to 'Transformacion'. Below it, there is a 'Show' dropdown set to '10' and a 'Search' input field. A table with 12 columns and 6 rows of data is displayed. The columns are: Id, NRRCR, HACR, ORLA, MAHO, CPBP, COAP, COHO, CECC, CYCC, CMCC, FLOS, and FRCC. The rows are numbered 1 to 6. Below the table, there is a 'Showing 1 to 6 of 6 entries' message and 'Previous', '1', and 'Next' navigation buttons. At the bottom, there is a 'Variables' section showing a list of variables and their data types. The variables are: Id, NRRCR, HACR, ORLA, MAHO, CPBP, COAP, COHO, CECC, CYCC, CMCC, FLOS, and FRCC. The data types are: character, factor, factor, factor, factor, factor, factor, factor, factor, factor, factor, factor, and factor.

	Id	NRRCR	HACR	ORLA	MAHO	CPBP	COAP	COHO	CECC	CYCC	CMCC	FLOS	FRCC
1	904-SAG	2	1	4	2	9	3	1	2	9	9	2	2
2	905-SAG	2	1	2	2	9	3	1	2	9	9	1	1
3	906-SAG	1	1	3	2	2	3	1	2	5	3	3	1
4	1229-ATR	1	1	3	1	6	4	4	2	5	5	2	1
5	1230-SAG	1	2	2	1	3	3	1	1	2	1	4	1
6	1231-SAG	1	2	3	1	4	3	1	1	5	1	4	3

Showing 1 to 6 of 6 entries

Previous 1 Next

Variables :
 Name "Id" [1] "NRRCR" [2] "HACR" [3] "ORLA" [4] "MAHO" [5] "CPBP" [6] "COAP" [7] "COHO" [8] "CECC" [9] "CYCC" [10] "CMCC" [11] "FLOS" [12]
 Class "character" "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor"
 Name "CNAHA" [14] "IPLH" [15] "COPE" [16] "CHEN" [17] "FOHO" [18] "SELB" [19] "PCPE_Ord" [20] "FLOR_Ord" [21] "Hum_ALPL" [22] "Hum_LOPE" [23] "Hum_ANAI" [24]
 Class "factor" "factor" "factor" "factor" "factor" "factor" "factor" "factor" "numeric" "numeric" "numeric"
 [26] [27] [28] [29] [30] [31] [32] [33] [34] [35]

Figura 2. Sección inicial de la aplicación.

The screenshot shows a software interface with a blue header bar containing 'Fichero', 'Análisis', and 'Conglomerados'. The 'Transformación' submenu is active, showing a 'View' tab and a 'Transform' tab. On the left, a 'Select a variable:' panel lists variables: CMCC, FLOS, FRCC, CNHA, IPLH, COPE, CNEN, FOHO, SELB, PCPE_Ord, FLOR_Ord, Num_ALPL, Num_LOPE, Num_ANAR, Num_ANHO, Num_LOHO, Num_NRCO, Num_NCOR, Num_DIIN, Num_RLAH, Num_RLHP, Num_LPHA, Num_NROH, Num_PMSF, Num_PMSC, Num_PMSL, and Num_PALM. Below this list are buttons for 'Numeric', 'Factor', and 'Eliminar'. On the right, a data table is displayed with 14 columns: Id, NRCR, HACR, ORLA, MAHO, CPBP, COAP, COHO, CECC, CYCC, CMCC, FLOS, and FRCC. The table contains 10 rows of data.

	Id	NRCR	HACR	ORLA	MAHO	CPBP	COAP	COHO	CECC	CYCC	CMCC	FLOS	FRCC
1	904-SAG	2	1	4	2	9	3	1	2	9	9	2	2
2	905-SAG	2	1	2	2	9	3	1	2	9	9	1	1
3	906-SAG	1	1	3	2	2	3	1	2	5	3	3	1
4	1229-ATR	1	1	3	1	6	4	4	2	5	5	2	1
5	1230-SAG	1	2	2	1	3	3	1	1	2	1	4	1
6	1231-SAG	1	2	3	1	4	3	1	1	5	1	4	3
7	1232-SP	1	3	2	1	7	1	3	2	5	3	3	3
8	1233-SAG	2	1	2	1	5	3	3	2	6	9	2	2
9	1234-SAG	2	2	2	1	6	2	1	2	7	6	2	2
10	1235-	2	3	2	1	5	3	1	2	8	7	1	1

Figura 3. Submenú “Transformación”.

Análisis por Factores

Los datos de entrada a este algoritmo deben ser variables continuas. El objetivo de este análisis es reducir la dimensionalidad de los datos al crear un nuevo conjunto de variables no correlacionadas a partir de un conjunto de variables correlacionadas, si las variables originales ya están no correlacionadas, entonces no se recomienda este análisis. Por tanto, primeramente, se calculan diferentes pruebas estadísticas: Prueba de Esfericidad de Bartlett, la Medida de Adecuación de la Muestra KMO (Keiser-Meyer-Olkin), así como la matriz de correlaciones y su determinante, con el objetivo de determinar si es posible realizar el análisis por factores en virtud de que las variables están correlacionadas entre sí y por tanto se pueden encontrar factores comunes subyacentes (Figura 4).

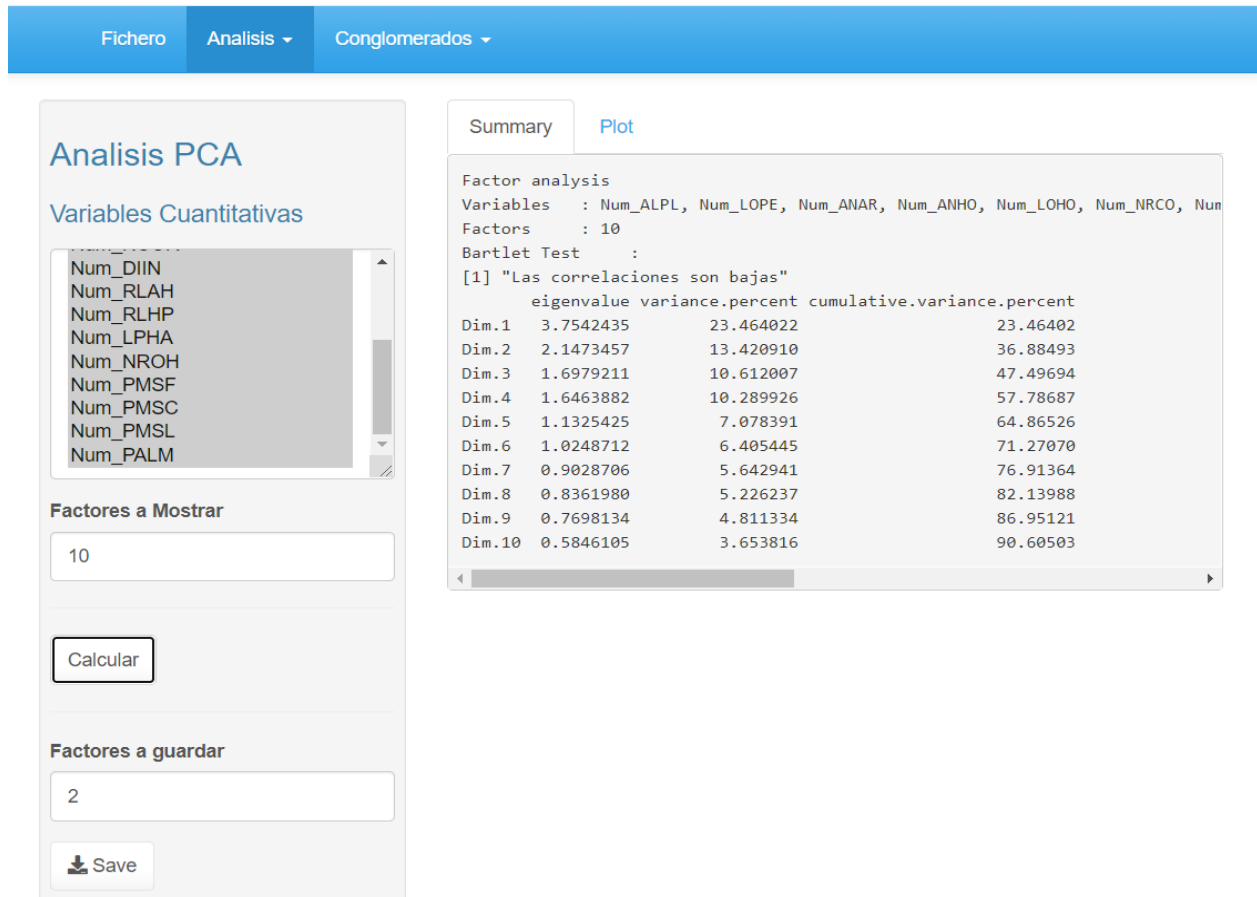


Figura 4. Resultados del análisis de componentes principales.

Una vez establecido el hecho de poder realizar este análisis, se procede a extraer el número óptimo de factores, para esto se calculan los autovalores, el porcentaje de varianza y el porcentaje de varianza acumulada y se muestra al usuario una gráfica donde se observan el número óptimo de factores con autovalor (varianza) mayor que uno, que son los establecidos por defecto para los cálculos posteriores. No obstante, el usuario puede desear mayor porcentaje de varianza acumulada y aumentar el número de factores (Figura 5).

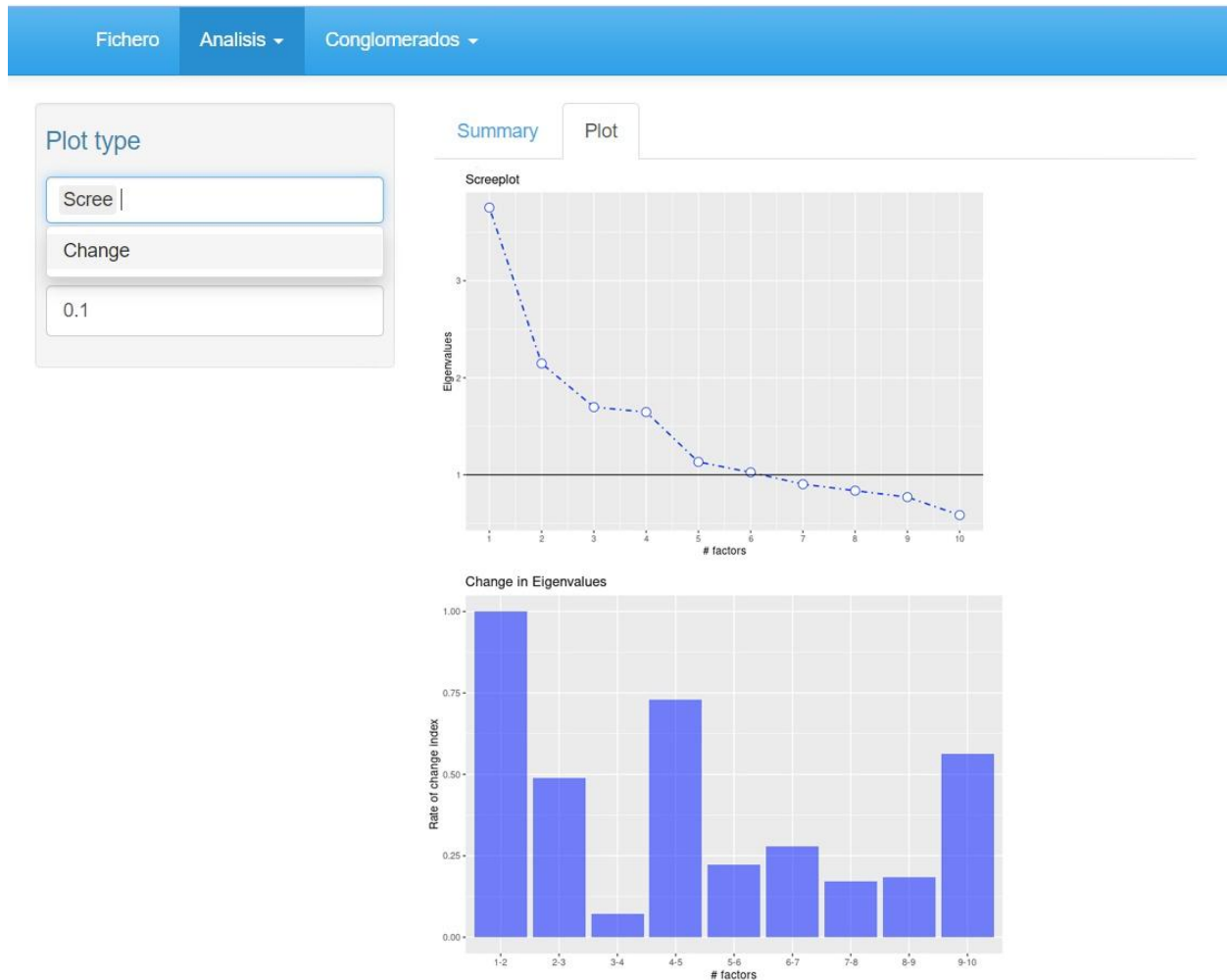


Figura 5. Gráficas del análisis de componentes principales.

Análisis de Correspondencias Múltiple

Para este análisis se pueden seleccionar variables cualitativas (ordinales y nominales). Inicialmente se calculan los autovalores, la varianza y la varianza acumulada para poder definir el número de componentes, luego se muestran los resultados en tablas acompañados de gráficos que ayudan al usuario al razonamiento estadístico y permiten observar la correspondencia entre las variables (Figura 6).



Figura 6. Resultados del análisis de correspondencia múltiple.

Métodos de Selección de atributos

Estos métodos de selección de atributos calculan un valor de importancia para cada atributo cualitativo y a la vez se muestran en orden descendiente. Para los atributos cuantitativos brinda una lista ordenada para cada método de selección, según la importancia de dichos atributos. El método mixto **HINoV.Mod**, presenta los atributos cuantitativos y cualitativos más importantes (Figura 7).

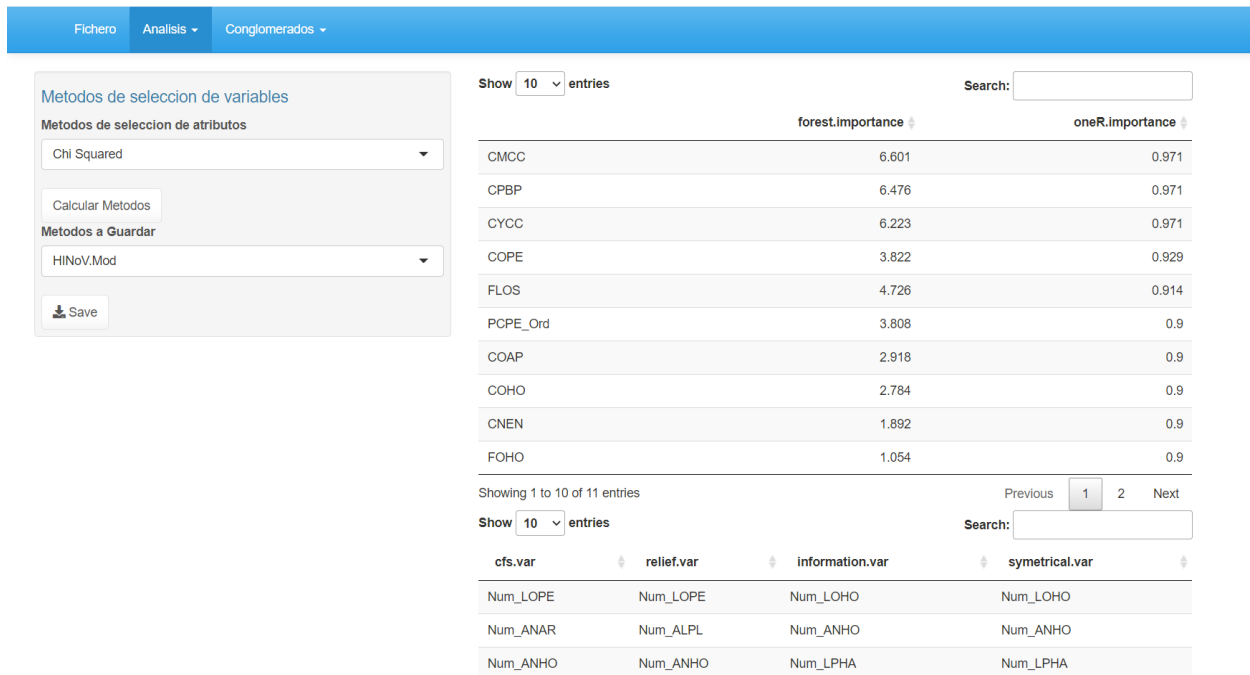


Figura 7. Resultados de los métodos de selección de atributos.

Análisis de Conglomerados

Este análisis incluye varios procedimientos clásicos de conglomeración y el usuario debe seleccionar la variante a efectuar combinando los métodos de cálculo de la matriz de distancia con los algoritmos de conglomeración que se ponen a su disposición. Se incluyen varios métodos de cálculo de la distancia: *Euclídea*, *Manhattan*, *Gower*, *Canberra* y *Minkowski*. La entrada de los datos debe provenir de tres resultados previamente obtenidos de los métodos de reducción de datos. Si los datos son de tipo ordinal se utilizará por defecto la métrica de *Gower*. Los métodos de conglomeración que se ponen a disposición del usuario son: varianza mínima (*Ward*), vecino más próximo (*Simple*), vecino más lejano (*Completo*), vinculación Inter-grupos (*Promedio*) y vinculación Intra-grupos (*Mcquitty*) que se encuentran en las bibliotecas del paquete *cluster*. Con la combinación de los datos de entrada al algoritmo, los métodos de cálculo de las distancias y los métodos de conglomeración se pueden generar varias clasificaciones para medir cuál de los métodos resulta más eficaz para el conjunto de datos en cuestión (Figura 8).



The image shows a web-based configuration interface for a cluster analysis tool. It features a light gray background with white input fields and blue buttons. At the top, there are two radio buttons: 'Selección de Variables' (unselected) and 'Componentes y Dimensiones Principales' (selected). Below these is a large blue button with a white play icon and the text 'Calcular modelo'. The interface is divided into several sections, each with a label in blue text: 'Dimensiones :', 'Factores :', 'Medida de distancia:', 'Metodo:', and 'Plot(s):'. Each section contains a white input field with a dropdown arrow on the right. The 'Dimensiones' field contains the number '10'. The 'Factores' field contains the number '6'. The 'Medida de distancia:' field contains the word 'gower'. The 'Metodo:' field contains 'ward.D'. The 'Plot(s):' field contains 'Dendogram'.

Figura 8. Entrada de datos, medidas de distancia, métodos de aglomeración y número de conglomerados.

Para facilitar la tarea de buscar la mejor combinación, se ofrecen coeficientes como el de Correlación Cofenético que puede ser usado para medir cuan bien la estructura jerárquica del dendrograma representa a las verdaderas distancias y el Aglomerativo que mide la estructura de aglomeración del conjunto de datos. Ambos coeficientes deben ser maximizados. Una vez seleccionado el método de cálculo de la distancia y el

método de conglomeración, se muestra el dendrograma y el próximo paso es definir el número de conglomerados a elegir, según la idea visual que se muestra y el cálculo del Índice de Dunn (Figura 9), el Ancho de Silueta (Figura 10).

```
Hierarchical cluster analysis
Variables      :
Method        : ward.D
Distance      : gower
Observations  : 70
Coeficiente Aglomerativo : 0.8762541
Indice de Dunn : 0.2069278
```

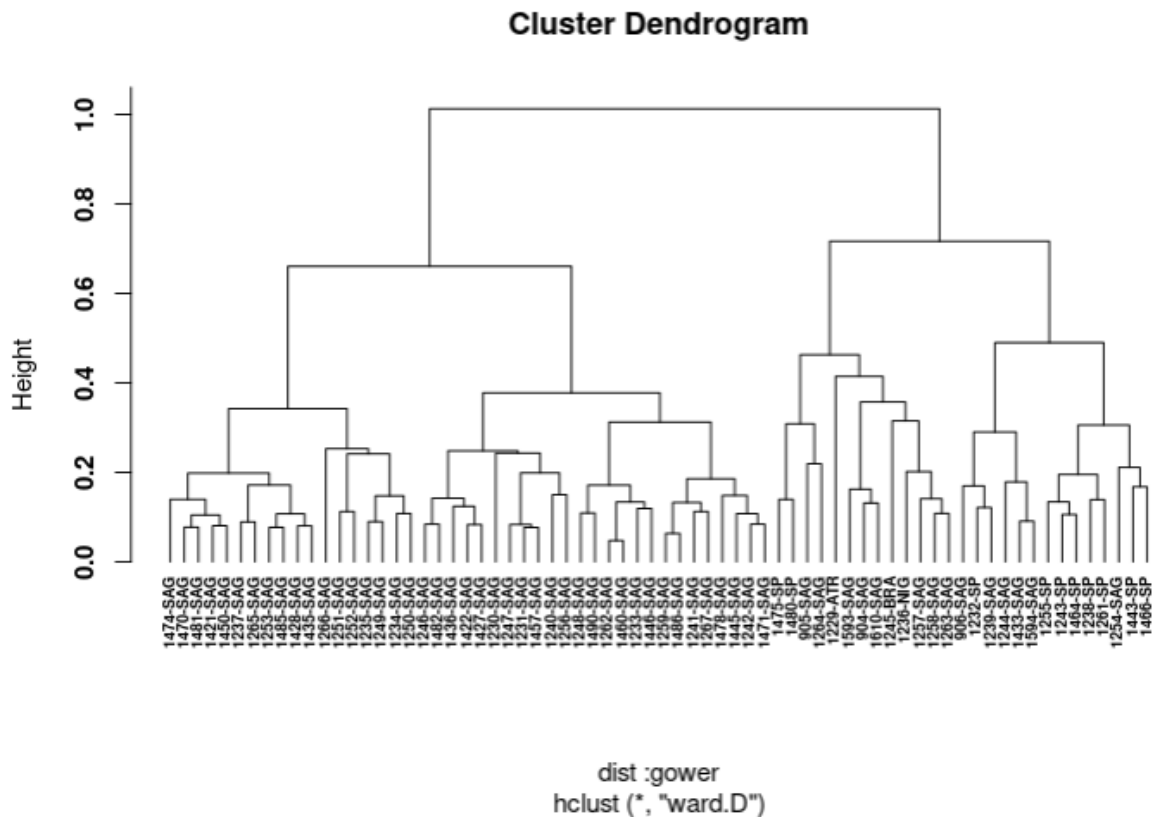


Figura 9. Resultados de análisis de *cluster*.

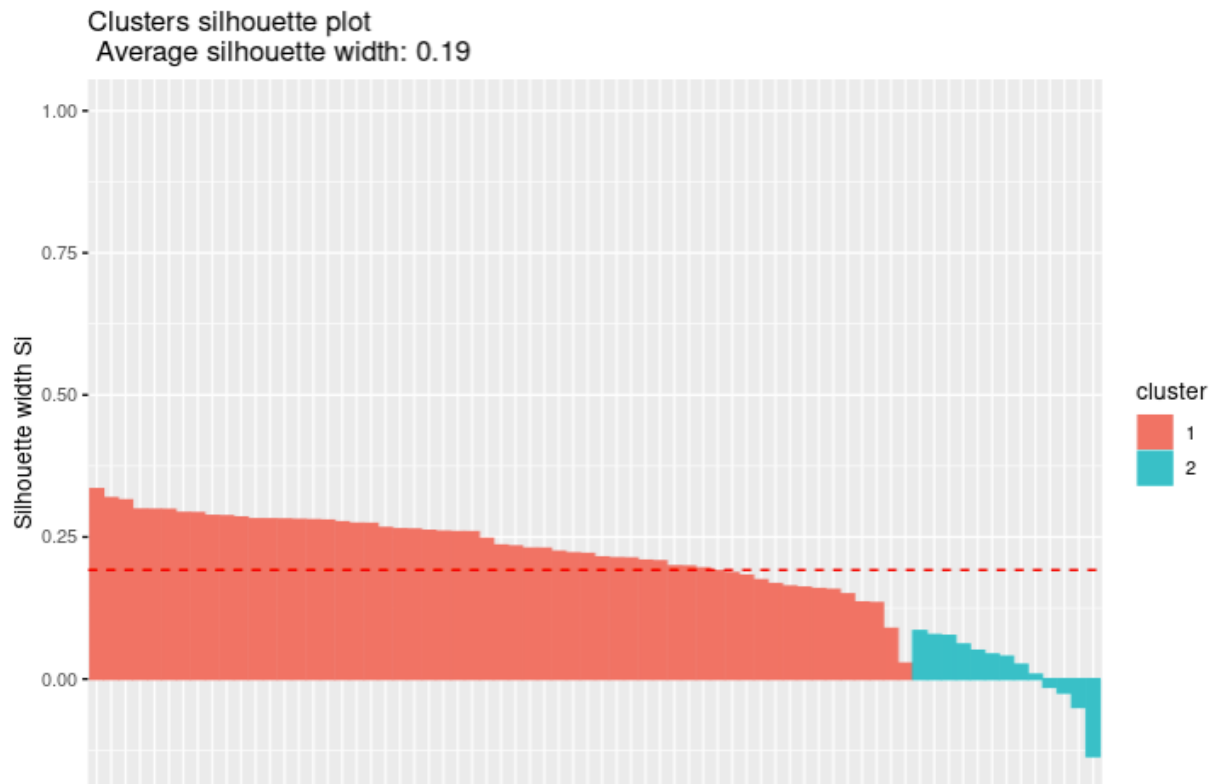


Figura 10. Índice de Silueta.

También se muestran como resultado la comparación entre el dendrograma calculado a partir de los datos tratados por los diferentes análisis con el de los datos originales permitiendo observar los diferentes cambios en la clasificación y la relación entre las accesiones (Figura 11).

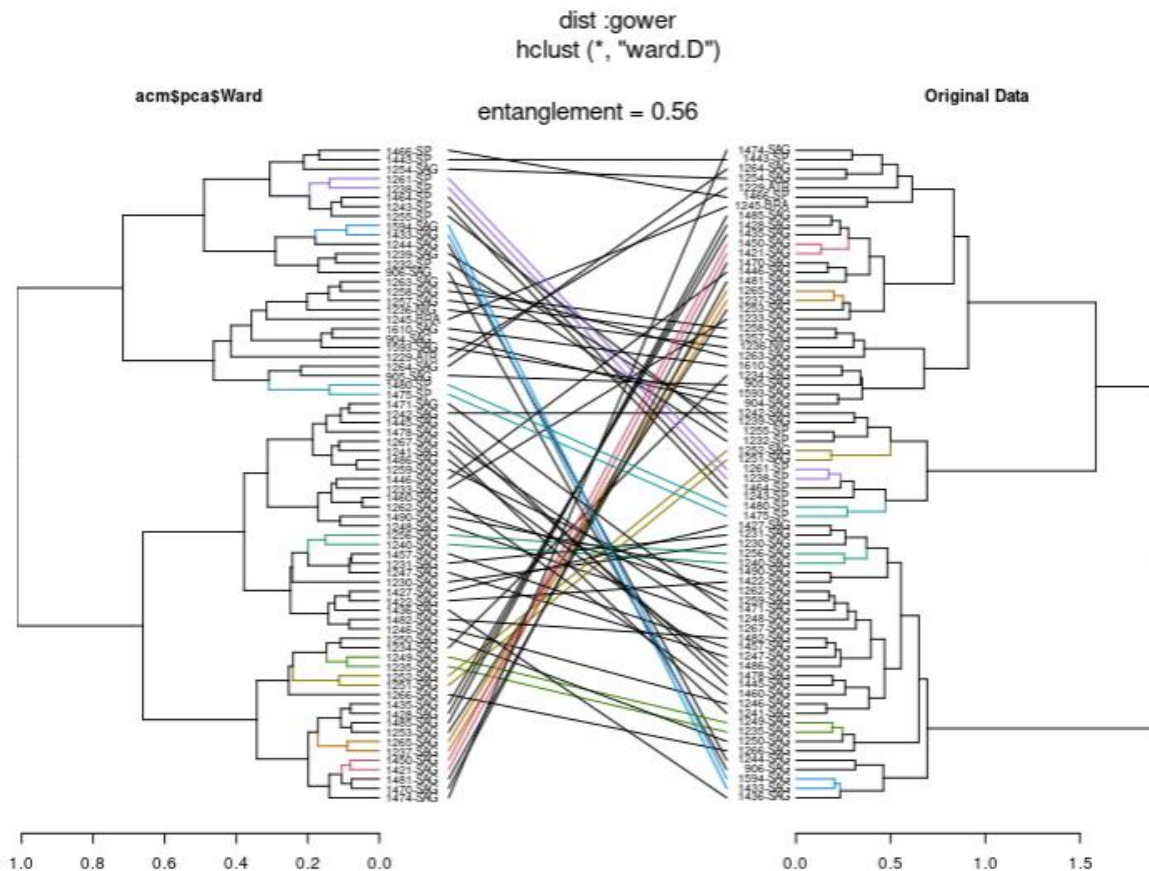


Figura 11. Resultados del análisis de *cluster*.

CONCLUSIONES

- Se obtuvo una aplicación *web* para la selección de descriptores en germoplasma vegetal a partir de métodos de selección de variables y de herramientas de inteligencia artificial y estadísticas usados en el *Software R*.
- La interfaz propuesta se puede acceder a través de la *web* sin tener que utilizar un ordenador que contenga *R*, ni dominar el *software* estadístico.

BIBLIOGRAFÍA

- ARNHOLT, A. T. 2018. Using a Shiny app to teach the concept of power. *Teaching Statistics*. ISSN 1467-9639. doi: 10.1111/test.12186.
- BEELEY, C. 2013. Web Application Development with R Using Shiny. Olton: Packt Publishing Ltd, first edition.
- BEELEY, C. and S.R. SUKHDEVE. 2018. Web Application Development with R Using Shiny. Birmingham UK: Packt Publishing.
- CISTERNA, A.; A. GONZÁLEZ-VIDAL; D. RUIZ; J. ORTIZ; A. GÓMEZ-PASCUAL; Z. CHEN and J.A. BOTÍA. 2021. PhenoExam: An R package and Web application for the examination of phenotypes linked to genes and gene sets. *bioRxiv*. <https://doi.org/10.1101/2021.06.29.450324>.

- CHANG, W.; J. CHENG; J.J. ALLAIRE; Y. XIE and J. MCPHERSON. 2020. Shiny: Web Application Framework for R. R package version 1.5. Retrieved Jul. 20, 2020. Available at CRAN.R-project.org/package=shiny.
- DWIVEDI, B. AND J. KOWALSKI. 2018. shinyGISPA: A web application for characterizing phenotype by gene sets using multiple omics data combinations. *PLoS ONE*, 13(2). ISSN 1932-6203. doi: 10.1371/journal.pone.0192563.
- LI, J.; B.CUI; Y. DAI; L. BAI and J. HUANG. 2018. BioInstaller: A comprehensive R package to construct interactive and reproducible biological data analysis applications based on the R platform. *PeerJ*, 6, ISSN 2167-8359. doi: 10.7717/peerj.5853.
- ELLIOTT, M.S. and L.M. ELLIOTT. 2020. Developing R Shiny Web Applications for Extension Education, Applied Economics Teaching Resources (AETR), *Agricultural and Applied Economics Association*, 2(4).
- LANDAU, S. and B. EVERITT. 2004. *A Handbook of Statistical Analyses Using SPSS*. Chapman & Hall/CRC, Boca Raton. ISBN 978-1-58488-369-2.
- LIHUA, J.; Y. WEN; J. YINGRU; L. YANG; W. ZHIZHAN; L. HAORAN; H. FANGFANG; L. JIAMING; C. TIANIAN and Z. HUIYONG. 2022. Development of interactive biological web applications with R/Shiny, *Briefings in Bioinformatics*, 23(1), <https://doi.org/10.1093/bib/bbab415>.
- MOLER, C. and MATHWORKS. 2012. MATLAB 8.0 and Statistics Toolbox 8.1. The MathWorks, Inc.
- NIJS, V. 2020. Radiant: Business Analytics using R and Shiny. R package version 1.3.2. Retrieved Jul. 21, 2020. Available at CRAN.R-project.org/package=radiant.
- R CORE TEAM. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at <https://www.R-project.org/>.
- SALAS, C. 2008. ¿Por qué comprar un programa estadístico si existe R? *Ecología Austral*, 18: 223-231.
- SATYAHADEWI, N. and H. PERDANA. 2021. Web Application Development for Inferential Statistics using R Shiny. Proceedings of the 1st International Conference on Mathematics and Mathematics Education (ICMMEd 2020).
- WEIGELT, P.; P. DENELLE; F. BRAMBACH and H. KREFT. 2022. BotanizeR: A flexible R package with Shiny app to practice plant identification for online teaching and beyond. *Plants People Planet*. 4(2): 122-127. <https://dx.doi.org/10.1002/ppp3.10226>.