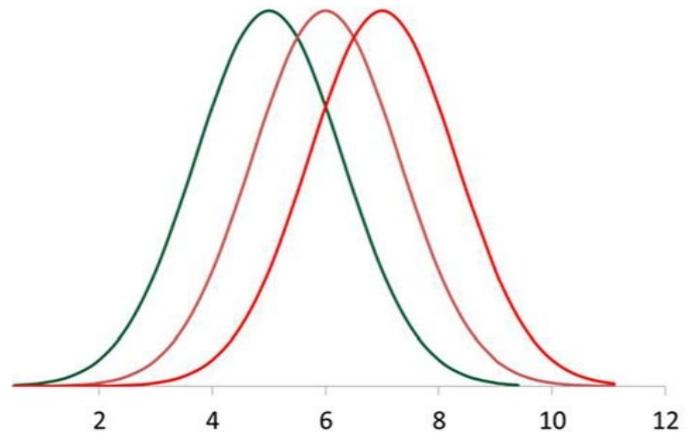


# Ejercicios de Mejora Genética

Grado en Ingeniería Forestal y del Medio Natural



Universidad de Oviedo



José Alberto Oliveira Prendes

Ana Belén Monteagudo Diz

# **Ejercicios de Mejora Genética**

**José Alberto Oliveira Prendes**

Dr. Ingeniero Agrónomo, Catedrático de Producción Vegetal,  
Departamento de Biología de Organismos y Sistemas,  
Universidad de Oviedo, Asturias.

**Ana Belén Monteagudo Diz**

Dra. en Biología,  
Asesoramiento para la conservación y regeneración de recursos naturales,  
actualmente colabora con la Consellería del Mar, Xunta de Galicia, Galicia.

13 de mayo de 2024

# ÍNDICE

	<b>Páginas</b>
<b>Análisis de la diversidad genética</b>	
1. Variación genética en poblaciones	3
2. Cuantificación de la variación genética	4
3. Introducción al GenAlEx 6.5	6
4. Introducción al NTSYS-pc 2.1	17
5. Cuantificación de las relaciones genéticas: Distancias genéticas y de similitud	23
6. Métodos de clasificación: Análisis Clúster	24
7. Métodos de ordenación: Análisis de Componentes Principales	31
8. Distancias geográficas	43
9. Test de Mantel	44
10. Bibliografía	49
<b>Análisis de pruebas genéticas</b>	
1. Pruebas genéticas	52
2. Introducción al programa SELEGEN-REML/BLUP	53
3. Análisis de ensayos de progenies de medios hermanos o de polinización abierta (plantas alógamas). Diseño en bloques completos al azar con una planta por parcela. Evaluación en una sola localidad	57
4. Análisis de ensayos de progenies de medios hermanos o de polinización abierta. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en una sola localidad	61
5. Análisis de ensayos de progenies de medios hermanos o de polinización abierta. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en varias localidades	70
6. Análisis de poblaciones o procedencias con progenies de medios hermanos dentro de la procedencia. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en una localidad	75
7. Análisis de ensayos con clones no emparentados. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en varias localidades. Estabilidad, Adaptabilidad y Productividad	78
8. Índice de selección	83
9. Agradecimientos	85
10. Bibliografía	85

# Análisis de la diversidad genética

## 1. Variación genética en poblaciones

Una población se define como el conjunto de seres vivos del mismo grupo o especie que se localizan en un área geográfica concreta. Si bien los seres vivos pertenecientes a una población son reconocibles como parte de la misma, no todos los individuos de la misma son idénticos, presentando diferencias en su morfología, función, comportamiento, etc.

Estas diferencias entre seres vivos vienen dadas por la diversidad de las frecuencias de los genotipos existentes, que da lugar a la aparición de variabilidad genética dentro de la población.

Aunque los miembros de una misma especie comparten los mismos genes, las diferencias genéticas se producen cuando distintos individuos albergan alelos diferentes. El conjunto de alelos existentes en una población constituye el *pool genético* que contiene toda la variabilidad fenotípica expresada en la población.

La frecuencia de ciertos genes puede verse alterada de una generación a otra generándose nuevos genotipos y contribuyendo a aumentar la variabilidad genética existente. Se podría hablar de tres fuentes principales de variabilidad genética:

- **Mutaciones:** Producidas por cambios en la información contenida en el material genético. Una mutación simple puede dar lugar a un cambio de gran efecto, sin embargo, el cambio evolutivo se basa en la acumulación de muchas mutaciones de efectos pequeños.
- **Flujo génico:** Producido por cualquier movimiento de material genético de una población a otra, como por ejemplo a través de las migraciones de las poblaciones o la introducción de un cultivo en una zona nueva.
- **Sobrecruzamiento:** Consistente en el intercambio de material genético durante la reproducción sexual entre dos cromosomas homólogos, que puede originar nuevas combinaciones genéticas en una población.

Cuando se habla de variabilidad poblacional también es importante tener en cuenta que el fenotipo de un individuo no sólo viene dado por el genotipo, sino que también tiene un importante componente ambiental, puesto que el medio ambiente puede influir significativamente en los organismos y cambiar la forma en que se expresan algunos genes. Por lo tanto, el genotipo dicta un fenotipo que se va a ver restringido según el entorno en el que se desarrolle, provocando variabilidad fenotípica en la población.

Desde los inicios de la Biología siempre se ha tratado de “clasificar” la variabilidad existente entre los seres vivos y delimitar grupos, estableciendo los límites de su variabilidad para situarlos en el conjunto de los seres vivos expresando su parentesco. La clasificación jerárquica de los seres vivos permite establecer de una forma sencilla las relaciones de parentesco entre los distintos grupos y así facilitar su estudio. En general, la clasificación se basa en la comparación de los caracteres que manifiesta cada unidad objeto de dicha clasificación. Los caracteres, son atributos de un organismo que pueden ser descritos, medidos, pesados, contados, etc. y pueden manifestar diferentes estados de carácter (por ejemplo, el color de ojos: castaños, azules, grises, verdes), permitiendo la caracterización de un ser vivo por el conjunto de los diferentes estados que expresan los caracteres estudiados.

Inicialmente, los estudios de caracterización se basaban en caracteres de observación en los individuos, es decir, en el fenotipo. Dentro de estos caracteres, se diferencia entre *caracteres cualitativos*, que describen cualitativamente al individuo (color del tallo, pubescencia, etc.) y *caracteres cuantitativos*, que muestran una distribución continua de fenotipos que puede ser cuantificada (peso, rendimiento, altura, etc.).

El desarrollo de las técnicas moleculares supuso un avance muy importante en los trabajos de caracterización, dado que no sólo se basan en los caracteres fenotípicos, sino que determinan las variaciones a nivel genotípico, permitiendo una clasificación mucho más exacta, puesto que dos individuos de igual fenotipo pueden contener diferencias a nivel genético, como por ejemplo la expresión isoenzimática diferente o el polimorfismo de un nucleótido en una determinada secuencia de ADN.

## **2. Cuantificación de la variación genética**

Los estudios de caracterización permiten cuantificar la variabilidad genética analizable en términos precisos mediante el empleo de diversos parámetros o índices estadísticos.

Los estadísticos más empleados para cuantificar la variabilidad intrapoblacional son:

- Porcentaje de loci polimórficos (P): Número de loci polimórficos entre el número total de loci. En general, se considera que un loci es polimórfico cuando el alelo más común tiene una frecuencia de menos de 0,95 (Cavalli-Sforza y Bodmer, 1981).
- Número de alelos ( $N_a$ ): Media del número de alelos que aparecen en los loci estudiados.
- Número efectivo de alelos ( $N_e$ ): Número de alelos que se esperaría en un locus en cada población ( $p_i$ = frecuencia del  $i$ -ésimo alelo en un locus).

$$N_e = \frac{1}{\sum p_i^2}$$

- Heterocigosidad observada (H<sub>o</sub>): Proporción de individuos heterocigotos observados.
- Heterocigosidad esperada (H<sub>e</sub>) o diversidad genética (D): Probabilidad de que en un locus cualquier par de alelos escogidos al azar sean diferentes entre sí.

$$H_e(j) = 1 - \sum p_i^2$$

Siendo: H<sub>e</sub> (j)= heterocigosidad esperada por locus  
 p<sub>i</sub> = frecuencia del i-ésimo alelo en un locus

La H<sub>e</sub> promedio de todos los loci es una estimación de la variabilidad genética de una población.

$$H_e = \sum_j^L h_j / L$$

Siendo: H<sub>e</sub> = heterocigosidad esperada promedio para varios loci  
 h<sub>j</sub> = heterocigosidad por locus  
 L= número total de loci

Sin embargo, si lo que se desea cuantificar es la variabilidad genética interpoblacional, los estadísticos más empleados son los *Estadísticos F (Wright)*. Conocidos como *Índices de fijación*, permiten el análisis de la estructura en poblaciones subdivididas y describen el nivel esperado de heterocigosidad en una población (Wright, 1950).

Los estadísticos F analizan los distintos niveles de la estructura poblacional diferenciándose entre:

**F<sub>ST</sub>** = Proporción de la varianza genética total contenida en una subpoblación en relación con la varianza genética total. Por lo tanto, indicaría el grado de diferenciación genética entre poblaciones en función de las frecuencias alélicas.

$$F_{ST} = 1 - (H_s / H_T)$$

El rango de diferenciación genética entre poblaciones ( $F_{ST}$ ) oscila entre 0 (no existe diferenciación genética) y 1 (alto grado de diferenciación genética entre poblaciones).

$F_{IS}$  = Coeficiente de endogamia dentro de poblaciones. Deficiencia o exceso de heterocigotos promedio en cada población. Los valores elevados de este indicarían un considerable grado de endogamia en la población.

$$F_{IS} = 1 - (H_I / H_S)$$

$F_{IT}$  = Coeficiente de endogamia total. Deficiencia o exceso de heterocigotos promedio en un grupo de poblaciones.

$$F_{IT} = 1 - (H_I / H_T)$$

Siendo:  $H_T$  = Diversidad genética total o la heterocigosidad esperada en la población total estimada a partir de las frecuencias alélicas combinadas.

$H_I$  = Diversidad genética dentro de una población o la heterocigosidad promedio observada en un grupo de poblaciones.

$H_S$  = Heterocigosidad promedio esperada estimada a partir de cada subpoblación.

### **3. Introducción al GenALEx 6.5**

La aplicación GenALEx está desarrollada en Visual Basic para Aplicaciones (VBA) para su uso en Excel (Peakall y Smouse, 2006; 2012). Está diseñada para la realización de análisis genéticos poblacionales empleando tanto datos codominantes como haploides o binarios.

Se diseñó inicialmente con fines educativos, como una herramienta para mejorar la comprensión de los análisis de genética poblacional y, dada su versatilidad y fácil manejo, se extendió también al ámbito científico-profesional.

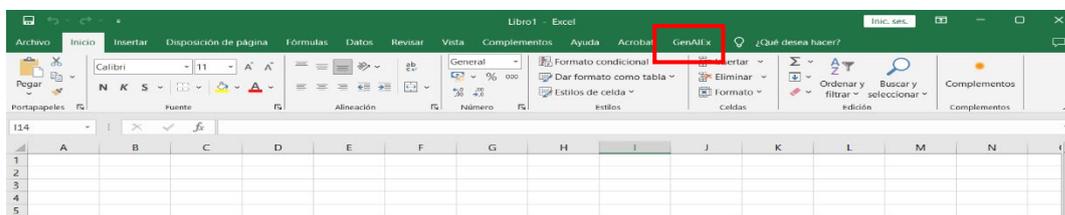
La aplicación se utiliza como un complemento de Excel, por lo que se puede configurar para que se cargue automáticamente cuando se inicializa Excel o bien se puede cargar sólo cuando se vaya a emplear.

Para emplear la aplicación debemos tener instalado en Excel el complemento

**Herramientas para análisis-VBA y habilitadas las macros** (configurar en el menú de

Excel *Archivo* → *Opciones de Excel* → *Complementos* - en el caso de las Herramientas VBA- o *Centro de Confianza* → *Configuración de macros* - en el caso de las macros).  
Descargada la aplicación de la página <https://biology-assets.anu.edu.au/GenAIEx/Download.html> debemos descomprimir la carpeta de archivos descargada (es recomendable tener una carpeta exclusiva para el programa, por ejemplo, una carpeta dentro de la carpeta PROGRAMAS creada en el directorio raíz C:).

Desde el menú de Excel *Archivo* → *Abrir* seleccionamos el archivo ejecutable de la aplicación C:/PROGRAMAS/GenAIEx6.5/**GenAIEx6.51b2** y de esta forma aparecerá la aplicación en el menú principal de Excel.



Si queremos que la aplicación se ejecute automáticamente al inicializar Excel, debemos seleccionar dentro del menú de GenAIEx la opción *Options* → *Install*. De lo contrario, abríamos el archivo ejecutable cada vez que vaya a utilizar la aplicación.

### 3.1. Preparación de los archivos de datos

Los datos analizados con GenAIEx deben tener formato numérico, por lo que si los datos están en formato texto (p ej. secuencias de ADN) debemos convertir esos datos a formato numérico y asegurarnos que las celdas de datos tengan formato “Número”.  
La matriz de datos a analizar debe localizarse en una hoja activa de Excel y presentar la estructura adecuada para que la aplicación descifre la información correctamente y la analice. Es posible diseñar la estructura de la matriz de datos a analizar mediante la opción dentro del menú GenAIEx *Create* → *Codominant/ Binary/ Haploid Template*, esta opción es adecuada para matrices con pocos datos; sin embargo, cuando disponemos de una matriz de datos grande la opción más rápida es estructurar nosotros los datos desde un inicio.

La estructura que deben seguir los datos sería la siguiente:

Nº de loci a analizar, en el ejemplo 2: GluA1 y GluA2

Nº de individuos totales, en este caso 898 individuos

Nº de poblaciones a analizar, en este caso 57

Nº de individuos analizados en cada una de las poblaciones: 17 individuos de la población 1, 25 individuos de la población 2 y así sucesivamente para las 57 poblaciones

Individuos analizados de la población 1

Individuos analizados de la población 2

Conjunto de datos genéticos

Nombre de cada población

Código numérico asignado a cada individuo

Para datos codominantes se utilizarían dos columnas por locus; en el caso de datos dominantes, haplotipos o secuencias codificadas se utilizaría una columna por locus

Primera fila:

Columna A: Nº de loci a analizar

Columna B: Nº de individuos totales a analizar

Columna C: Nº de poblaciones a analizar

Columna D en adelante: Nº de individuos analizados en cada población

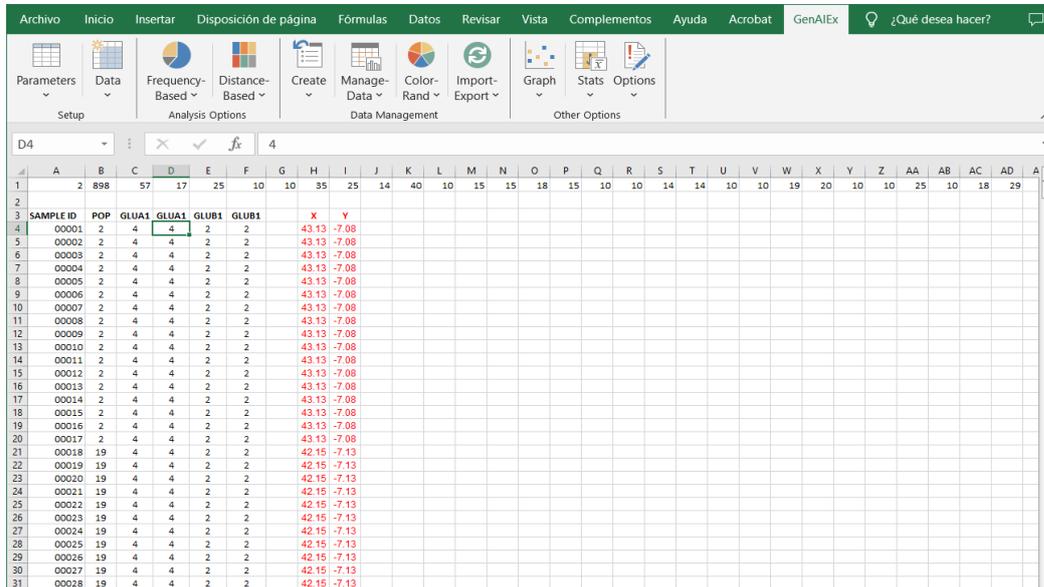
Segunda fila: Pondremos un título identificativo de los datos o bien dejamos en blanco.

Tercera fila: Identificativo de los datos que introduciremos.

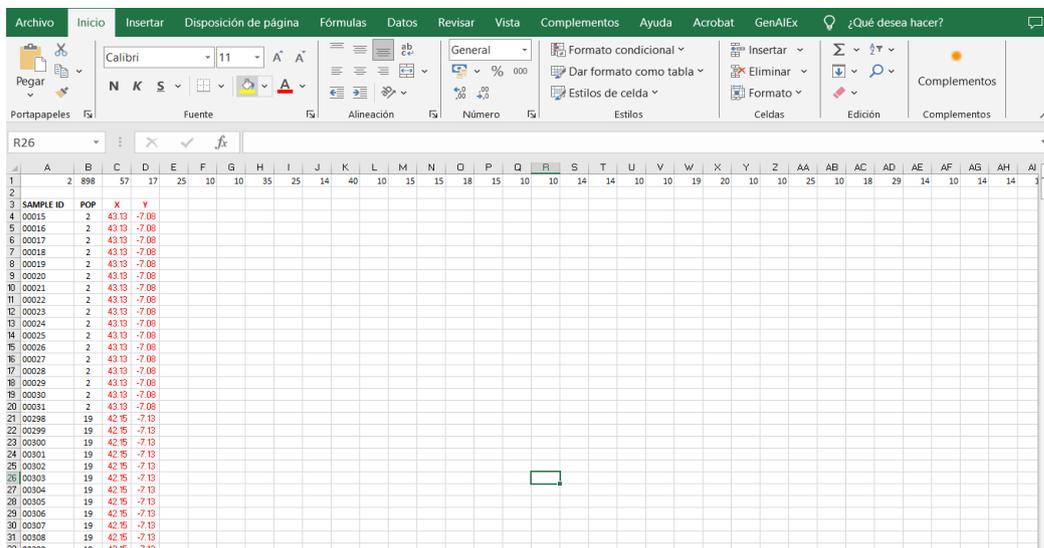
Cuarta fila en adelante: Listado de los datos por individuo.

En caso de disponer de las coordenadas geográficas y que estas fuesen de interés para el análisis, podrían:

- Introducirse los datos en las columnas a continuación de los datos genéticos y separados de estos en una columna.



- O bien en una hoja Excel distinta, conservando la información de la fila 1 de la hoja con la matriz de datos genéticos.



Las coordenadas pueden introducirse con números decimales o enteros y en formato UTM o Latitud/longitud.

### 3.2. Codificación de datos

Como se ha mencionado en el apartado anterior, los datos deben introducirse en formato numérico y con una o dos columnas por locus, dependiendo de si se trata de datos dominantes, haploides/haplotipos o codominantes. En muchas ocasiones, la matriz de datos de la que disponemos no cumple ese criterio, polo que es necesario realizar una codificación de los datos para adaptarlos al formato del GenAIEx.

Utilizaremos como ejemplo un archivo Excel con los datos del análisis de caracterización de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.) mediante el estudio de las subunidades de gluteninas de alto peso molecular.

En el archivo Gluteinas\_datos.xlsx, hoja **Resultados\_Gluteinas** (carpeta Ejercicio 1) disponemos de los datos crudos del análisis de los alelos de los loci *Glu-A1* y *Glu-B1* y la codificación aplicada en este caso.

Para el locus *Glu-A1* se analizaron los alelos *Glu-A1a*, *Glu-A1c*, *Glu-A1b*, *Glu-A1* y *Glu-A1bb*, a los que se le asigna un número del 1 al 5. En el caso del locus *Glu-B1* se analizaron los alelos *Glu-B1e*, *Glu-B1f*, *Glu-B1cf*, *Glu-B1at*, *Glu-B1as*, *Glu-B1ba*, *Glu-B1bx*, *Glu-B1ci*, *Glu-B1cg* y *Glu-B1ch*, a los que se les asigna un número del 1 al 10. Así, en formato GenAIEx el genotipo del individuo estará representado por el n° asignado a los alelos presentes en dicho individuo.

SAMPLE ID	ECOTIPO	Código del alelo					Código del alelo										GLUB1			
		1	2	3	4	5	1	2	3	4	5	6	7	8	9	10				
00001	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00002	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00003	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00004	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00005	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00006	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00007	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00008	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00009	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00010	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00011	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00012	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00013	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00014	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00015	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00016	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00017	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00018	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00019	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00020	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00021	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00022	2	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00023	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00024	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00025	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00026	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00027	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00028	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00029	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00030	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00031	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00032	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00033	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2
00034	19	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2

Datos crudos de Gluteinas y codificación de genotipos para análisis con el programa GenAEx

SAMPLE ID	ECOTIPO	Código del alelo																			
		1	2	3	4	5	6	7	8	9	10										
00055	61	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	2
00056	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00057	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00058	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00059	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00060	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00061	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00062	61	0	0	0	0	1	5	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00063	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00064	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00065	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00066	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00067	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00068	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00069	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00070	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00071	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00072	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00073	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00074	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00075	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00076	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00077	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00078	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00079	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00080	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00081	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00082	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00083	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00084	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00085	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00086	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00087	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2
00088	91	0	0	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	2

En el mismo archivo Gluteinas\_datos.xlsx se encuentran la matriz de datos ya transformada en la hoja **Datos\_formato análisis** y la matriz de datos geográfico en la hoja **Coordenadas**.

Datos con formato para programa GenAEx

SAMPLE ID	ECOTIPO	GLUA1	GLUA1	GLUB1	GLUB1
00001	2	4	4	2	2
00002	2	4	4	2	2
00003	2	4	4	2	2
00004	2	4	4	2	2
00005	2	4	4	2	2
00006	2	4	4	2	2
00007	2	4	4	2	2
00008	2	4	4	2	2
00009	2	4	4	2	2
00010	2	4	4	2	2
00011	2	4	4	2	2
00012	2	4	4	2	2
00013	2	4	4	2	2
00014	2	4	4	2	2
00015	2	4	4	2	2
00016	2	4	4	2	2
00017	2	4	4	2	2
00018	19	4	4	2	2
00019	19	4	4	2	2
00020	19	4	4	2	2
00021	19	4	4	2	2
00022	19	4	4	2	2
00023	19	4	4	2	2
00024	19	4	4	2	2
00025	19	4	4	2	2
00026	19	4	4	2	2
00027	19	4	4	2	2
00028	19	4	4	2	2
00029	19	4	4	2	2
00030	19	4	4	2	2
00031	19	4	4	2	2
00032	19	4	4	2	2
00033	19	4	4	2	2
00034	19	4	4	2	2
00035	19	4	4	2	2
00036	19	4	4	2	2

### 3.3. Estudio de la variabilidad genética poblacional

La aplicación GenAEx permite realizar múltiples análisis para la caracterización genética de las poblaciones y establecer cómo se estructura esa variación en la población y en relación con el conjunto de poblaciones.

En este caso, vamos a emplear la aplicación para determinar los principales parámetros que se utilizan para el análisis de la variación genética intra e interpoblacional.

#### **Ejercicio 1. Análisis de la variabilidad genética.**

**Ficheros: Gluteinas\_analisis.xlsx, Gluteinas\_analisis\_resultados.xlsx**

##### **Objetivo:**

Se realizó la caracterización de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.) mediante el estudio de las subunidades de gluteninas de alto peso molecular utilizando electroforesis SDS-PAGE, obteniéndose los genotipos por individuo para los alelos de los loci *Glu-A1* y *Glu-B1*.

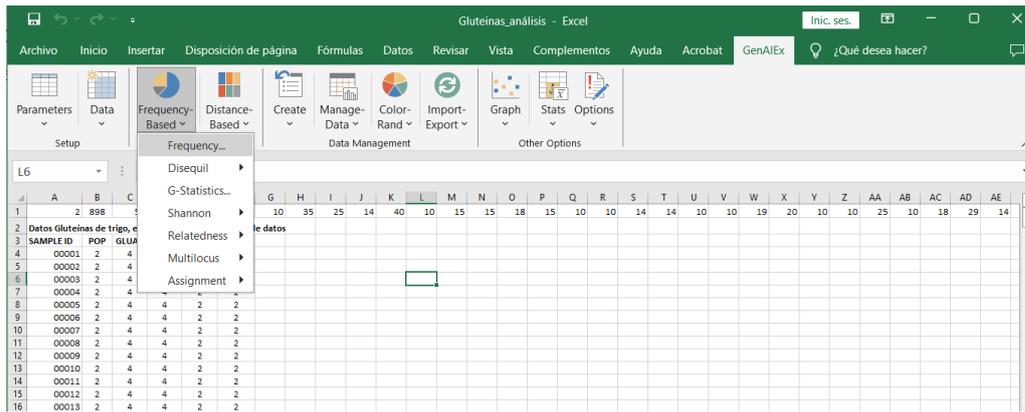
##### **Cuestiones:**

Calcular los estadísticos para cuantificar la variabilidad intrapoblacional: porcentaje de loci polimórficos (P), número de alelos ( $N_a$ ), número efectivo de alelos ( $N_e$ ), heterocigosidad observada ( $H_o$ ) y esperada ( $H_e$ ), así como los estadísticos F de variabilidad interpoblacional.

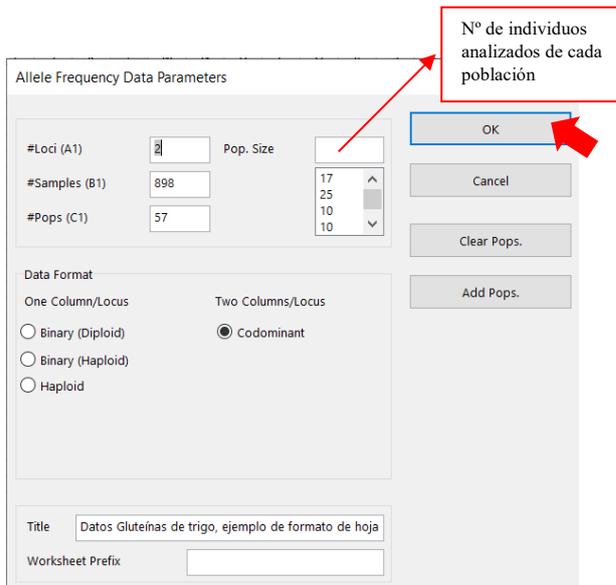
#### **Resolución Ejercicio 1.**

- Abrir Excel y abrir el archivo ejecutable GenAEx6.51b2.
- Abrir el archivo Gluteinas\_analisis.xlsx que contiene los datos genotípicos de los individuos analizados en cada ecotipo.
- Accedemos al menú de la aplicación GenAEx y seleccionamos la opción *Frequency-Based* → *Frequency...*

Nota: En el archivo Gluteinas\_analisis.xlsx sólo existe una hoja de datos, pero si el archivo de análisis tuviese más de una hoja, debemos asegurarnos antes de seleccionar la opción de análisis en el GenAEx que la hoja activa es la hoja que contiene la matriz de datos a analizar.



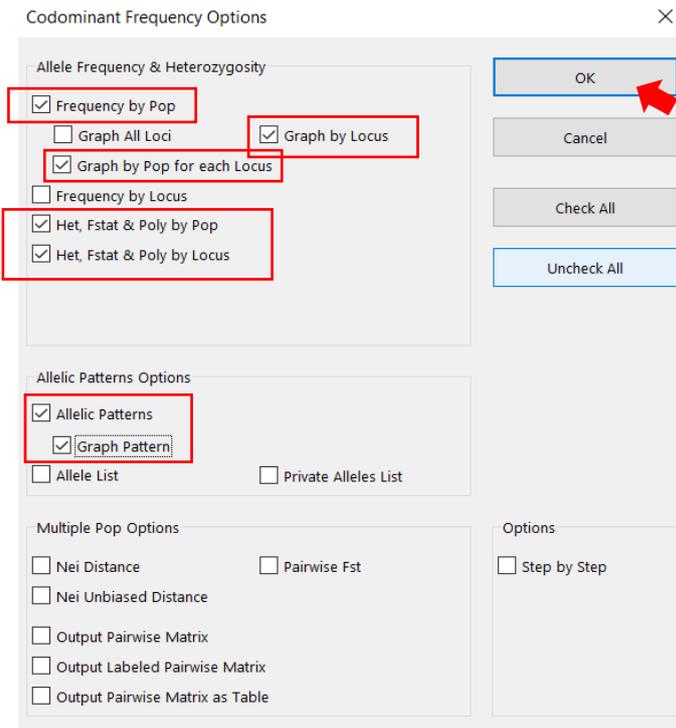
Se abrirá una primera pantalla en la que aparece el resumen de nuestra matriz de datos y en la que podremos detectar si existe algún error de lectura de la matriz que tengamos que subsanar. Si está todo correcto clicaremos en *OK*.



A continuación, se abrirá una nueva pantalla en la que podremos seleccionar los parámetros del análisis:

*Allele Frequency & Heterozygosity*: Cálculo de las frecuencias por locus y por locus en cada población y heterocigosidades y índices F por población y por locus analizado.

*Allele Patterns*: Patrón alélico obtenido en cada población con la identificación de los alelos comunes o únicos de cada población.



Los resultados obtenidos del análisis los encontraremos en varias hojas generadas por la aplicación.

Locus	Allele/n	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9	Pop10	Pop11	P
GLUA1	N	17	25	10	10	35	25	14	40	10	15	15	
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.275	0.000	0.000	0.000	0.000
	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	3	0.000	0.000	0.000	0.000	0.000	0.080	0.000	0.000	0.725	0.000	0.000	0.000
	4	1.000	1.000	1.000	0.000	1.000	0.920	0.000	0.571	0.000	1.000	0.267	1.000
GLUB1	N	17	25	10	10	35	25	14	40	10	15	15	
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.071	1.000	0.000	0.000	0.000	0.000
	2	1.000	0.920	1.000	1.000	1.000	1.000	0.929	0.000	1.000	1.000	1.000	1.000
	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	9	0.000	0.080	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

En las cuatro primeras hojas se listan los resultados obtenidos del cálculo de las frecuencias genéticas (AFP) y los gráficos correspondientes a las frecuencias por locus totales (AGF) y por locus por población (GLUA1 AGP, GLUB1 AGP).

En las **Hojas HFP y HFL** se encuentran los parámetros de variabilidad genética calculados por población (HFP) y por locus (HFL). Al inicio de la hoja se observa la tabla con los resultados para los parámetros: nº de alelos ( $N_a$ ), nº efectivo de alelos ( $N_e$ ), índice de información equivalente al índice de Shannon-Weaver ( $I$ ), heterocigosidad observada ( $H_o$ ) y esperada ( $H_e$ ), heterocigosidad esperada corregida ( $uHe$ -Unbiased Expected Heterozygosity- heterocigosidad esperada calculada corregida en función del tamaño muestral), e índice de fijación ( $F$ ). A continuación, encontraremos los resultados de los índices  $F$  ( $F$ -Statistics) y el porcentaje de loci polimórficos.

Finalizado el listado datos, encontraremos el código de nomenclatura de las abreviaturas empleadas para parámetro y la fórmula empleada para su cálculo.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
370	Pop51		50.00%														
371	Pop52		50.00%														
372	Pop53		100.00%														
373	Pop54		50.00%														
374	Pop55		50.00%														
375	Pop56		0.00%														
376	Pop57		0.00%														
377																	
378	Mean		33.33%														
379	SE		4.40%														
380																	
381	Na = No. of Different Alleles																
382	Ne = No. of Effective Alleles = 1 / (Sum pi^2)																
383	I = Shannon's Information Index = -1 * Sum (pi * Ln (pi))																
384	Ho = Observed Heterozygosity = No. of Hets / N																
385	He = Expected Heterozygosity = 1 - Sum pi^2																
386	uHe = Unbiased Expected Heterozygosity = (2N / (2N-1)) * He																
387	F = Fixation Index = (He - Ho) / He = 1 - (Ho / He)																
388	Where pi is the frequency of the ith allele for the population & Sum pi^2 is the sum of the squared population allele frequencies.																
389																	
390	Fs = (Mean He - Mean Ho) / Mean He																
391	Ft = (Ht - Mean Ho) / Ht																
392	Fst = (Ht - Mean He) / Ht																
393	Nm = [(1 / Fst) - 1] / 4																
394	Key: Mean He = Average He across the populations. Mean Ho = Average Ho across the populations.																
395	Ht = Total Expected Heterozygosity = 1 - Sum tpi^2 where tpi is the frequency of the ith allele for the total & Sum tpi^2 is the sum of the squared total allele frequencies.																
396																	
397	Tip: Choose the Step by Step option to show Mean He, Mean Ho and Ht values.																
398																	
399	Note:																
400	Mean F-Statistics represent arithmetic averages. See G-Statistics for an alternative mean Fst estimate, calculated based on the average Hs and Ht over loci.																
401																	
402																	
403																	
404																	

En la **Hoja APT** se encuentran los resultados del patrón alélico obtenido para cada población con el número de alelos que son únicos a cada población y número de alelos comunes con las otras poblaciones analizadas.

- En el archivo `Gluteinas_analisis_resultados.xlsx` se encuentran los resultados obtenidos en los análisis para la comprobación de la correcta realización del ejercicio. Dados los datos obtenidos en el cálculo de las frecuencias alélicas (Hojas AFP, AGF, GLUA1 AGP y GLUB1 AGP) se observaba que existen alelos con una frecuencia muy

superior al resto, puesto que para el locus *Glu-A1* el alelo predominante en todos los ecotipos es el alelo denominado 4 (*Glu-A1y*) y en el caso del locus *Glu-B1* sería el alelo denominado 2 (*Glu-B1f*).

En la Hoja APT donde se encuentra el resultado del patrón alélico, se observa la heterocigosidad esperada ( $H_e$ ) del global de loci por ecotipo, observando que existen ecotipos monomórficos para ambos loci ( $H_e=0$ ) y ecotipos muy polimórficos, con valores de  $H_e>0,30$ . Si observamos los datos de heterocigosidad por población y locus (Hoja HFP) se observa la existencia de ecotipos que son polimórficos para un solo locus, con una  $H_e=0$  en un locus y una  $H_e>0$  para el otro locus (por ejemplo la población 2, equivalente al ecotipo 19) y ecotipos que son polimórficos para ambos loci analizados (por ejemplo la población 7, equivalente al ecotipo 95).

La heterocigosidad observada ( $H_o$ ) fue 0 en todos los loci y en todos los individuos analizados, indicando que todos eran homocigotos.

Los datos resumen por locus obtenidos en la Hoja HFL, muestran que el locus *Glu-A1* presenta mayor variabilidad que el *Glu-B1*, con un mayor número efectivo de alelos ( $N_e=1,311$ ) y mayor heterocigosidad esperada ( $H_e=0,178$ ).

Mean and SE over Pops for each Locus			
		GLUA1	GLUB1
<b>N</b>	Mean	15.754	15.754
	SE	0.868	0.868
<b>Na</b>	Mean	1.526	1.193
	SE	0.083	0.053
<b>Ne</b>	Mean	1.311	1.077
	SE	0.051	0.027
<b>I</b>	Mean	0.269	0.079
	SE	0.040	0.024
<b>Ho</b>	Mean	0.000	0.000
	SE	0.000	0.000
<b>He</b>	Mean	0.178	0.050
	SE	0.027	0.016
<b>uHe</b>	Mean	0.184	0.051
	SE	0.028	0.016
<b>F</b>	Mean	1.000	1.000
	SE	0.000	0.000

En cuanto a la estructura de la diversidad genética de la población, tanto en la Hoja HFP como en la HFL se obtienen los resultados de los estadísticos F de Wright. Se observa que el valor para ambos coeficientes de endogamia  $F_{is}$  (dentro de poblaciones) y  $F_{it}$  (total) toman el valor 1, esto de debería que a los individuos analizados eran homocigotos con una  $H_o=0$ . Por otra parte, el valor de  $F_{st}>0,25$  en ambos loci indicaría

la existencia de una gran diferenciación genética entre los ecotipos de trigo para los loci analizados.

F-Statistics and Estimates of Nm over All Pops for each Locus					
All Pops.	Locus	Fis	Fit	Fst	
	GLUA1	1.000	1.000	0.569	
	GLUB1	1.000	1.000	0.504	
	Mean	1.000	1.000	0.537	
	SE	0.000	0.000	0.032	

#### 4. Introducción al NTSYS-pc 2.1

El NTSYSpc está compuesto por un conjunto de subprogramas estadísticos o módulos para la elaboración de análisis multivariantes. Cada módulo puede ser utilizado por separado o bien realizar un análisis más completo con el uso secuencial de dos o más módulos.

El programa se desarrolló inicialmente para la realización de análisis taxonómicos en biología, de ahí el nombre del programa (Numerical Taxonomy SYSTEM for personal computer) (Rohlf, 1998). Sin embargo, se ha utilizado en otras disciplinas como la ecología, ingeniería, humanidades etc.

Su uso más común es para la obtención de matrices de similitud y distancias entre pares de objetos y representar esta información mediante análisis de agrupación de objetos similares, como es el caso del análisis clúster, o bien representando espacialmente los objetos en uno o más ejes (análisis componentes principales). Por ello, es un programa de gran utilidad para el estudio de la variabilidad genética y la representación de la estructura genética de las poblaciones y la realización de análisis comparativos entre las mismas.

A pesar de la aparición de nuevos softwares de análisis, el NTSYSpc continúa siendo utilizado por su versatilidad y sencillez, empleando archivos directamente de Excel (.XLS, formato Excel 93-2003) se pueden analizar tanto datos moleculares como datos de caracterización cuantitativos y cualitativos. En el paquete NTSYS-pc 2.1 se incluye el editor NTedit, que permite guardar los datos en formato .XLS en formato “.NTS” para su uso en los análisis a realizar en los módulos del NTSYSpc, sin necesidad de realizar una adaptación compleja previa de la matriz de datos para su análisis.

#### 4.1. Preparación de los archivos de datos

El programa NTSYSpc trabaja con matrices de datos que deben tener una estructura determinada para que el programa interprete correctamente la información que se le suministra, así la matriz de datos debe ir precedida de una *cabecera* en dónde se le indicará al programa:

- Tipo de matriz de datos que estamos introduciendo (1 = Matriz rectangular, 2 = Matriz de distancias, 3 = Matriz de similitud, 4 = Matriz diagonal)
- Número de filas
- Número de columnas

En el caso de las matrices de datos introducidas mediante archivos “.txt” debemos indicar si las filas/columnas de datos tienen etiquetas de identificación (por ejemplo, el nombre de los individuos evaluados, nombre del carácter evaluado). Para ello indicaríamos con una “L” que la etiqueta está entre la cabecera y la matriz de datos, con una “B” si la etiqueta está al inicio de cada fila/columna y con una “E” si la etiqueta está al final de cada fila/columna. Para un mejor funcionamiento del programa, las etiquetas no den tener más de 16 caracteres.

- Datos perdidos. Es posible trabajar con matrices en las que falta algún dato. Para identificar esos datos de los que se carece debe emplearse una codificación diferente a la empleada para el resto de los datos e indicarle al programa cuál es esa codificación en la cabecera. Generalmente, los datos perdidos suelen indicarse con un 999.

En el siguiente ejemplo se observa un archivo “.txt” en cuya cabecera se indicaría que se trata de una matriz rectangular (1) con 62 filas en las que la etiqueta de identificación de cada individuo analizado está al principio de cada fila (62B), 15 columnas en las que la etiqueta con el nombre de cada carácter se encuentra entre la cabecera y la matriz de datos (15L) y los datos perdidos, de haberlos, se indicarían con un 999.

1	57B	15L	999												
	Glu-A1a	Glu-A1c	Glu-A1b	Glu-A1y	Glu-A1bb		Glu-B1e	Glu-B1f	Glu-B1cf		Glu-B1at		Glu-B1as		Glu-B1ba
1	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.05	0.95	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	1.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00
4	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	1.00	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00

Es posible introducir los datos a partir de un archivo Excel, en cuyo caso no sería necesario indicar la presencia de etiquetas de filas y columnas. Los archivos deben

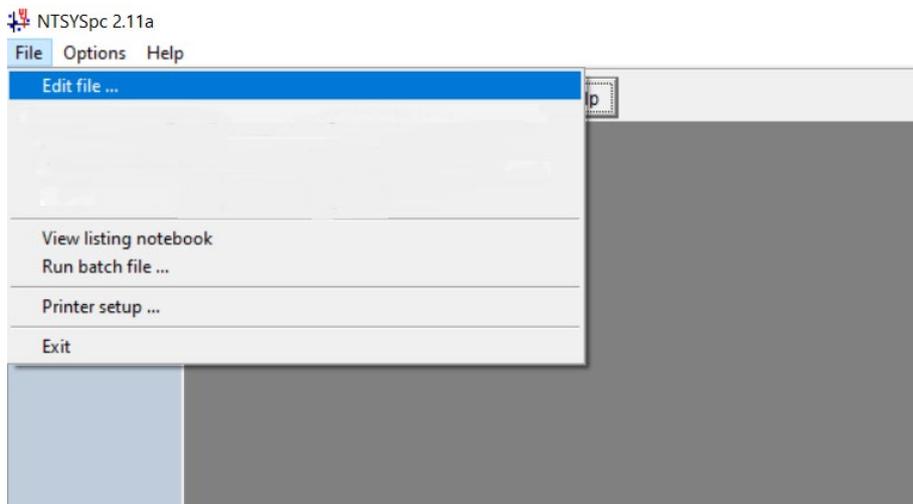
guardarse en con extensión .xls formato Excel 93-2003 para que sean reconocidos por el NTSYSpc.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	1	57	15	999													
2		Glu-A1a	Glu-A1c	Glu-A1b	Glu-A1y	Glu-A1bb	Glu-B1e	Glu-B1f	Glu-B1cf	Glu-B1at	Glu-B1as	Glu-B1ba	Glu-B1bx	Glu-B1ci	Glu-B1cg	Glu-B1ch	
3	2	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	19	0.00	0.00	0.00	1.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
5	21	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	61	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	91	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	92	0.00	0.00	0.08	0.92	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	95	0.00	0.00	0.00	0.57	0.43	0.07	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	97	0.28	0.00	0.72	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	99	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	102	0.00	0.00	0.00	0.27	0.73	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	107	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	109	0.00	0.00	0.00	1.00	0.00	0.00	0.67	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	112	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

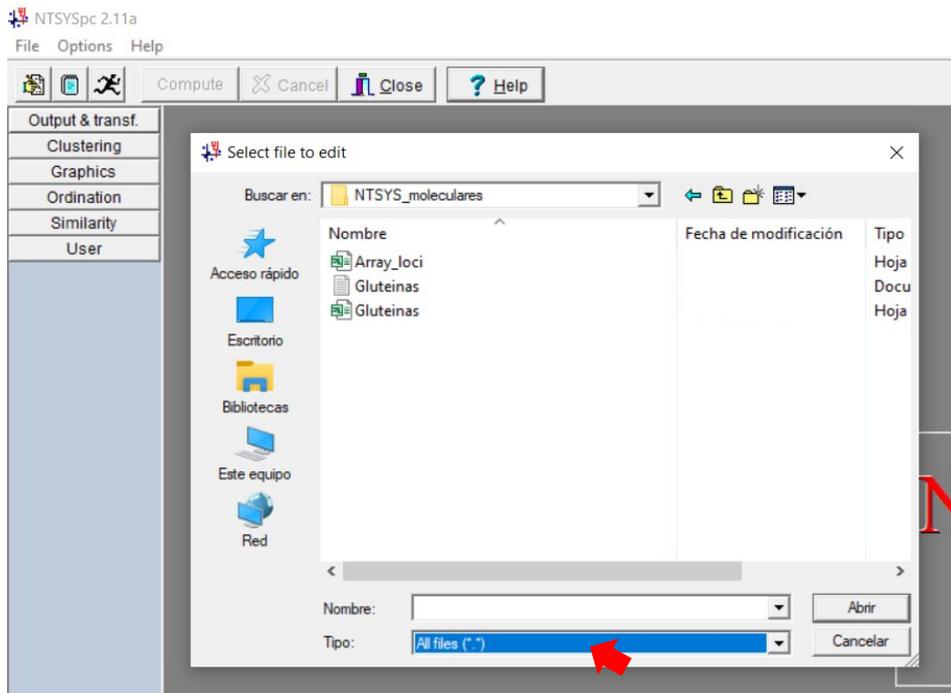
#### 4.2. Introducción de la matriz de datos en el NTSYSpc

Una vez preparada la matriz de datos en formato “.txt” o “.xls” debemos abrirla mediante el editor del NTSYSpc y guardarla en formato.NTS para poder trabajar con ella.

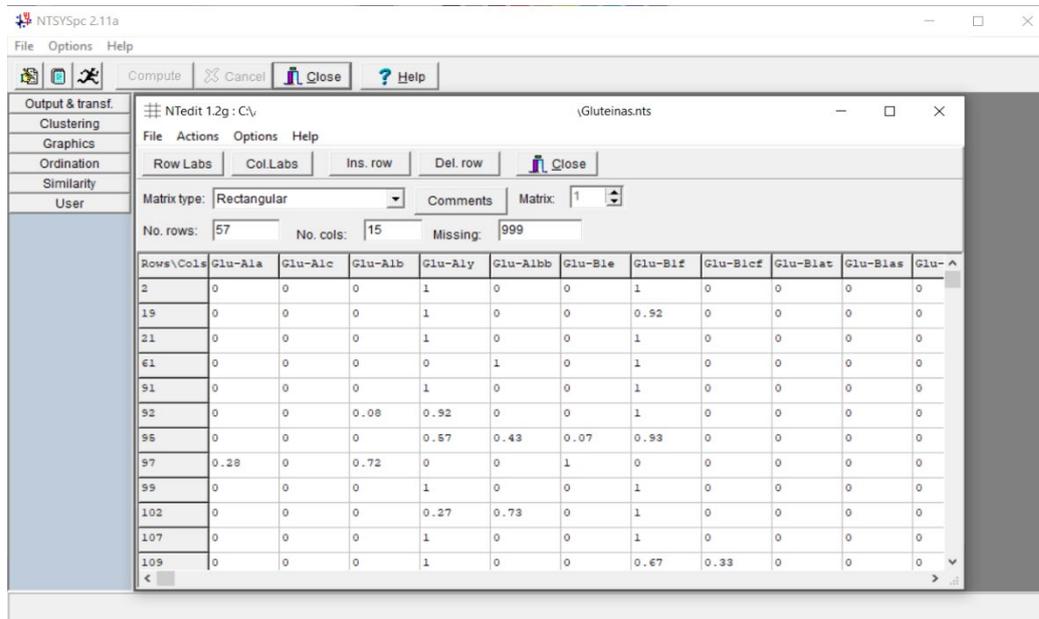
Para ello abrimos el programa y vamos al menú situado en la esquina superior izquierda *File* → *Edit file*



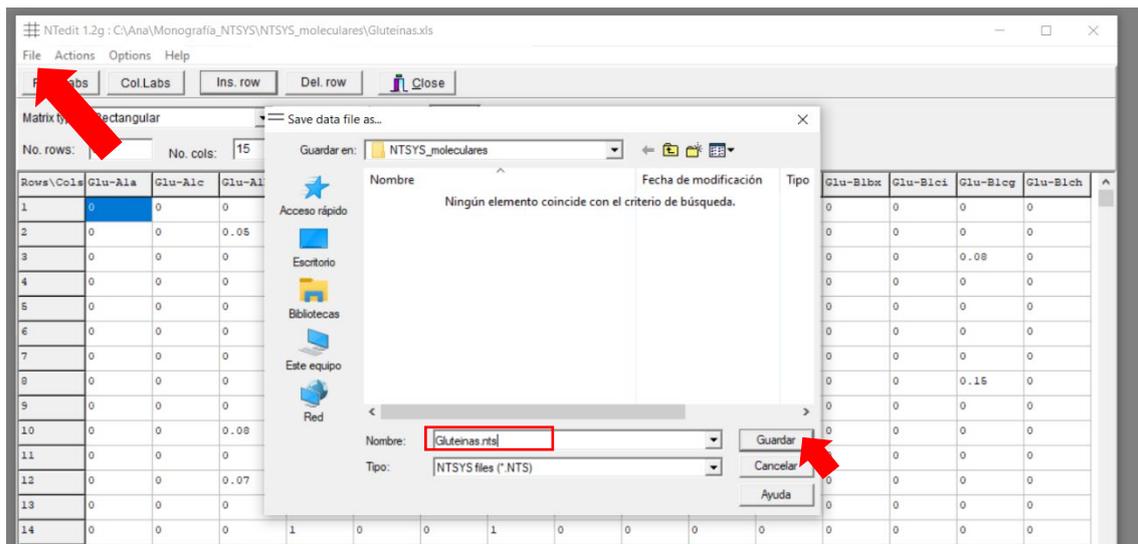
Seleccionamos el archivo con el que deseamos trabajar, teniendo en cuenta que en *Tipo de archivo* debemos tener seleccionado *All files (\*.\*)* para que podamos visualizar los archivos a introducir.



Abierto el archivo, comprobamos que el programa haya interpretado bien los datos de la cabecera y la matriz de datos esté correcta.



A continuación, en el menú del NEdit *File*, seleccionamos *Save file as*. Nombramos al archivo con la denominación que queramos e incluimos al final del nombre del archivo la extensión “.nts”.



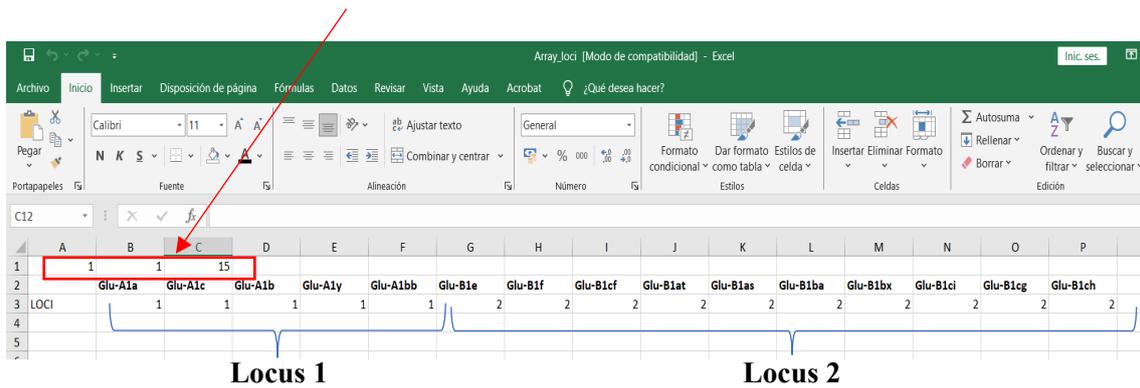
De esta forma ya tendríamos de la matriz de datos en el formato adecuado para la realización de los análisis deseados.

#### 4.3. Matrices secundarias

En ocasiones puede ser necesario hacer “aclaraciones” sobre la matriz de datos principal para que nuestro análisis sea más específico.

En las matrices de datos moleculares se suele disponer de datos de más de un loci, con datos de varios alelos de cada loci. En este caso debemos indicar al programa los datos de qué columnas pertenecen a qué locus en concreto y de cuántos loci diferentes disponemos de datos. Para ello, se realizará una matriz de datos, denominada como **array**, en los que figure en la primera fila la cabecera pertinente, en segunda fila los nombres de los alelos de los que se dispone datos, en el mismo orden en que aparecen en la matriz de datos base y en la tercera fila nombraremos a qué locus pertenece cada alelo denominando los loci con números desde el 1 en adelante.

Tipo de matriz = 1    N° filas = 1    N° columnas = 15



En el ejemplo se indica que en la matriz de datos base tendremos datos de dos loci, donde las primeras 5 columnas pertenecen a datos de 5 alelos del locus 1 y las siguientes 10 columnas pertenecen a datos de 10 alelos del locus 2.

También es posible que en la matriz de datos base tengamos datos de varios individuos de poblaciones distintas.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Poblacion	Ind	Glu-A1a	Glu-A1c	Glu-A1b	Glu-A1y	Glu-A1bb	Glu-B1e	Glu-B1f	Glu-B1f	Glu-B1at	Glu-B1as	Glu-B1ba	Glu-B1bx	Glu-B1ci	Glu-B1cg	Glu-B1ch
2	Pb1	1	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Pb1	2	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	Pb1	3	0.00	0.00	0.05	0.95	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	Pb1	4	0.00	0.00	0.43	0.57	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	Pb1	5	0.00	0.00	0.1	0.7	0.2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	Pb1	6	0.00	0.00	0.43	0.57	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	Pb1	7	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	Pb1	8	0.00	0.00	0.06	0.94	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	Pb1	9	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	Pb1	10	0.00	0.00	0.00	1.00	0.00	0.00	0.78	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	Pb2	11	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	Pb2	12	0.00	0.00	0.00	1.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	Pb2	13	0.00	0.00	0.06	0.94	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	Pb2	14	0.00	0.00	0.07	0.93	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	Pb2	15	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	Pb2	16	0.00	0.00	0.1	0.9	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	Pb2	17	0.00	0.00	0.00	1.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	Pb2	18	0.00	0.00	0.14	0.86	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	Pb2	19	0.00	0.00	0.00	1.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	Pb2	20	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	Pb3	21	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	Pb3	22	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	Pb3	23	0.00	0.00	0.00	0.32	0.68	0.36	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	Pb3	24	0.00	0.00	0.00	0.93	0.07	0.24	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
26	Pb3	25	0.06	0.00	0.00	0.25	0.69	0.12	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27	Pb3	26	0.11	0.00	0.00	0.67	0.22	0.11	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
28	Pb3	27	0.00	0.00	0.1	0.9	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29	Pb3	28	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	Pb3	29	0.00	0.00	0.00	1.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
31	Pb3	30	0.00	0.00	0.00	1.00	0.00	0.1	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
32																	

Si deseamos realizar los análisis tanto dentro de la población como entre poblaciones tendremos que indicarle al programa qué filas de datos pertenecen a cada población, mediante una matriz **ID**, en la que en la primera fila indicaríamos en la cabecera el tipo de matriz (1), el número de filas (n° de individuos totales analizados) y el número de columnas (en este caso sería una única columna donde se indica mediante numeración, de 1 al n° máximo de poblaciones del archivo, a qué población pertenece cada fila de datos).

Tipo matriz = 1 N° filas = 30 (30 individuos analizados) N° columnas = 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	1	30	1												
2	1	1													
3	2	1													
4	3	1													
5	4	1													
6	5	1													
7	6	1													
8	7	1													
9	8	1													
10	9	1													
11	10	1													
12	11	2													
13	12	2													
14	13	2													
15	14	2													
16	15	2													
17	16	2													
18	17	2													
19	18	2													
20	19	2													
21	20	2													
22	21	3													
23	22	3													
24	23	3													
25	24	3													
26	25	3													
27	26	3													
28	27	3													
29	28	3													
30	29	3													
31	30	3													
32															

## 5. Cuantificación de las relaciones genéticas: Distancias genéticas y de similitud

Seleccionados los caracteres o variables a considerar para la caracterización, cada uno de los individuos sujetos al análisis estará representado por los valores que tomen estas variables en cada uno de ellos. Este es el punto de partida para realizar la clasificación de los individuos, clasificación en la que se deberá determinar lo similares o disimilares (distantes) que son entre sí, en función de lo diferentes que resulten ser sus representaciones en el espacio de las variables.

Existen multitud de índices de similitud (similaridad) y coeficientes de distancia o disimilaridad con propiedades y utilidades distintas, por lo que su utilización dependerá del tipo de análisis que deseemos realizar (Soto *et al.*, 2006). En general, los índices de distancia o similitud se basan en la distancia (considerando a los individuos como vectores en el espacio de las variables) o en coeficientes de correlación.

Los índices de similitud se recomienda emplearlos para comparar individuos mediante caracteres con escala nominal o datos de doble estado (presencia/ausencia), obteniendo un valor que varía entre el 0 y el 1, siendo el 1 el de máxima similitud. Los índices de similitud más empleados son:

- Coeficiente de coincidencia simple (Simple matching coefficient -SM\*) (Sokal y Michener, 1958)
- Índice de Jaccard (J\*) (1908)
- Índice de Rogers y Tanimoto (RT\*) (1960)
- Índice de Dice (1945) o Sorensen (1948) (DICE\*)

Los coeficientes de distancia o disimilaridad representan la similitud como la proximidad de los individuos con respecto a los demás y se recomiendan para analizar datos cuantitativos, moleculares o mixtos (cuantitativos, cualitativos y/o moleculares). Se obtiene una matriz simétrica con valores que varían desde 0 a  $\infty$ , siendo el 0 el valor de máxima similitud. Los coeficientes de distancia más utilizados son:

- Datos cualitativos: Manhattan (MANHAT\*), Distancia Euclídea (EUCLID\*), Distancia Euclídea al cuadrado (EUCLIDSQ\*).
  - Datos moleculares: Distancia de Nei (NEI72\*) (1972), Hillis (1984), Prevosti (Wright, 1978) Cavalli-Sforza y Edwards o distancia del arco (ARC\*) (1967) o Rogers (1972).
  - Datos mixtos: Distancia de Gower (1971)
- (\* nomenclatura del NTSYSpc para dichos índices y coeficientes)

## 6. Métodos de clasificación: Análisis Clúster

Los métodos de clasificación permiten la búsqueda de grupos (*clústers*) similares lo más homogéneos posible para clasificar los individuos de un estudio y realizan una ordenación jerárquica de los grupos establecidos. Los grupos son *a priori* desconocidos y el objetivo con el análisis clúster es encontrar un conjunto de grupos a los que ir asignando los distintos individuos por algún criterio de homogeneidad, de forma que los individuos dentro de cada grupo sean lo más similares entre sí que sea posible y siendo los grupos entre ellos tan diferentes como sea posible. Por lo tanto, se hace imprescindible definir una medida de similitud o bien de distancia para ir clasificando a los individuos en unos u otros grupos e ir construyendo un **árbol** o **dendrograma** en función del grado de semejanza (o diferencia) que existe entre ellos.

Los métodos de construcción de dendrogramas pueden ser aglomerativos (van uniendo los individuos progresivamente), jerárquicos (presentan distintos rangos o niveles de unión de los individuos) y secuenciales (repiten el mismo proceso repetidas ocasiones hasta unir todos los individuos), siendo estos últimos los más frecuentes. Como métodos secuenciales más utilizados están:

- Encadenamiento simple (SINGLE): llamado también método del vecino más próximo (*nearest neighbour*), o de los mínimos, se unen los individuos por el valor de mayor semejanza o menor distancia.
- Encadenamiento completo (COMPLETE): llamado también método del vecino más lejano (*furthest neighbour*) o de máximos, se unen los individuos por el valor menor de semejanza o mayor de distancia.
- Encadenamiento por la media aritmética (UPGMA): se obtiene la media de los coeficientes de semejanza (o diferencia) entre individuos y se unen según este valor. En este caso, cada vez que se obtiene un grupo se ha de volver a calcular los valores medios. **Este es el método más utilizado para la construcción de dendrogramas.**
- Encadenamiento por la media ponderada (WPGMA): En el caso del encadenamiento UPGMA se promedian los valores originales mientras en el encadenamiento WPGMA, se pondera el cálculo de las distancias según sea el orden en que fueron fusionados los grupos.
- Método de Ward: utiliza como criterio el aumento del valor de la suma de las distancias al cuadrado de cada individuo al centroide, como resultado de la unión de dos grupos.

## **Ejercicio 2. Análisis clúster.**

**Ficheros: Gluteinas.xls, Array\_loci.xls**

### **Objetivo:**

Realizada la caracterización de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.) mediante el estudio de las subunidades de gluteninas de alto peso molecular se dispone de las frecuencias alélicas para los alelos de los loci *Glu-A1* y *Glu-B1*.

### **Cuestiones:**

Realizar el análisis de agrupación y obtener el dendrograma de agrupación de los ecotipos estudiados.

### **Resolución Ejercicio 2.**

- Introducir la matriz de datos base de frecuencias alélicas (Gluteinas.xls) y la matriz con los datos de los loci (Array\_loci.xls) en el NTSYSpc a través del editor NTedit para guardar los archivos en formato “.nts” (ver apartado 4.2. *Introducción de la matriz de datos en el NTSYSpc*).
- Calcular la matriz de distancias genéticas entre individuos

Para calcular una matriz de similitud con datos genéticos se emplea el módulo

*Similarity* → *Genetic distance*

Introducimos los datos para el análisis:

Input data file= Matriz base de frecuencias alélicas Gluteinas.nts

By rows?= seleccionamos esta casilla ya que los datos de los individuos están en filas.

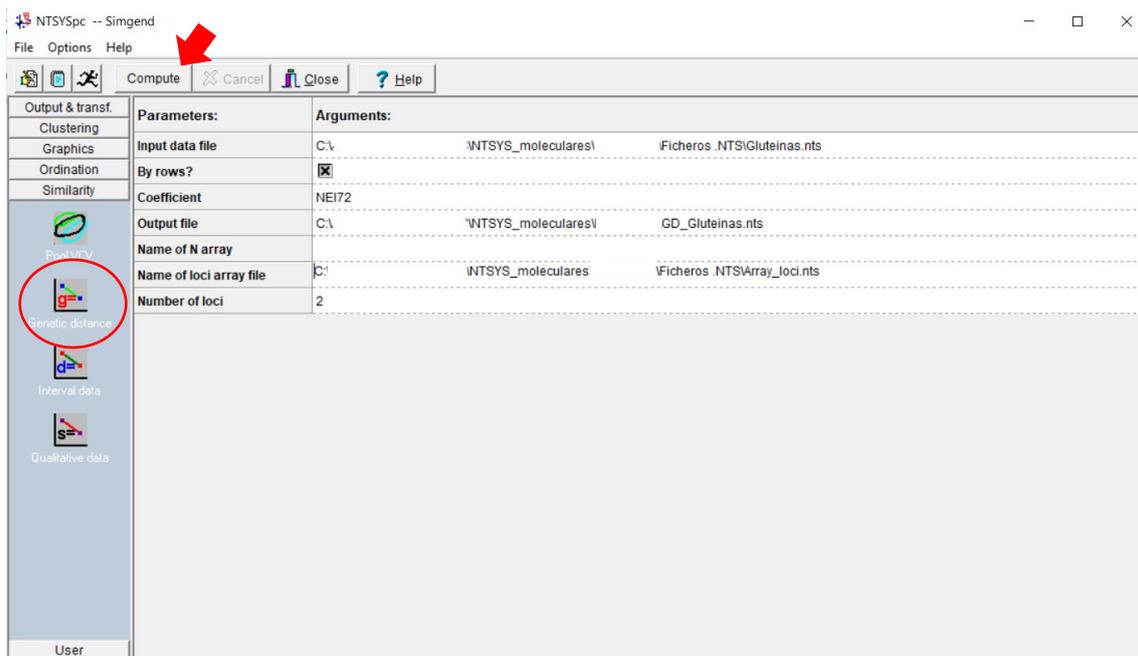
Coefficient= seleccionamos NEI72 para el cálculo de la distancia genética mediante la distancia de Nei (1972).

Output file= nombre y ubicación del archivo que se generará tras el análisis con la matriz de distancias genéticas, en este caso GD\_Gluteinas.nts.

Name of loci array file= introducimos la matriz con el nombre de los loci y la correspondencia de los alelos Array\_loci.nts

Number of loci= 2, nº de loci a analizar, que en este caso son dos (*Glu-A1* y *Glu-B1*).

Introducidos los archivos y datos necesarios, pulsamos la tecla *Compute*.

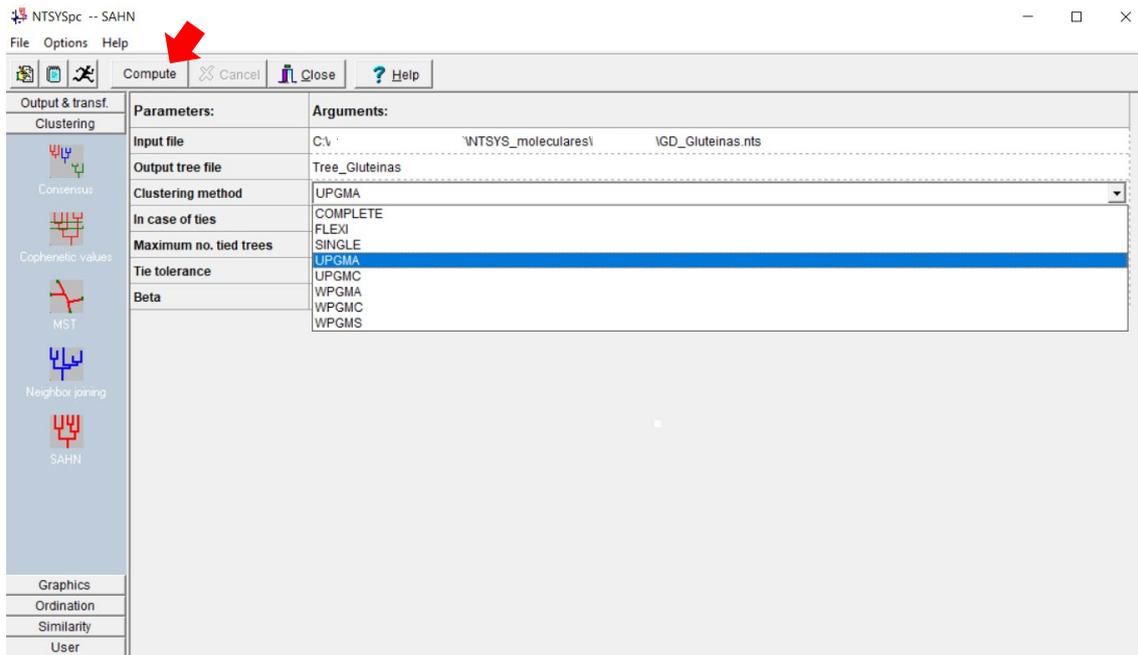


Terminado el análisis se abrirá una ventana con el *Reporting list*, en el cual se hace un resumen del procedimiento de análisis que se está ejecutando.

- Construir el dendrograma a partir de la matriz de distancias genéticas

GD\_Gluteinas.nts

Seleccionamos el módulo *Clustering* → *SAHN*



Introducimos los datos para la construcción del dendrograma:

Input file= archivo con la matriz de distancias genéticas GD\_Gluteinas.nts

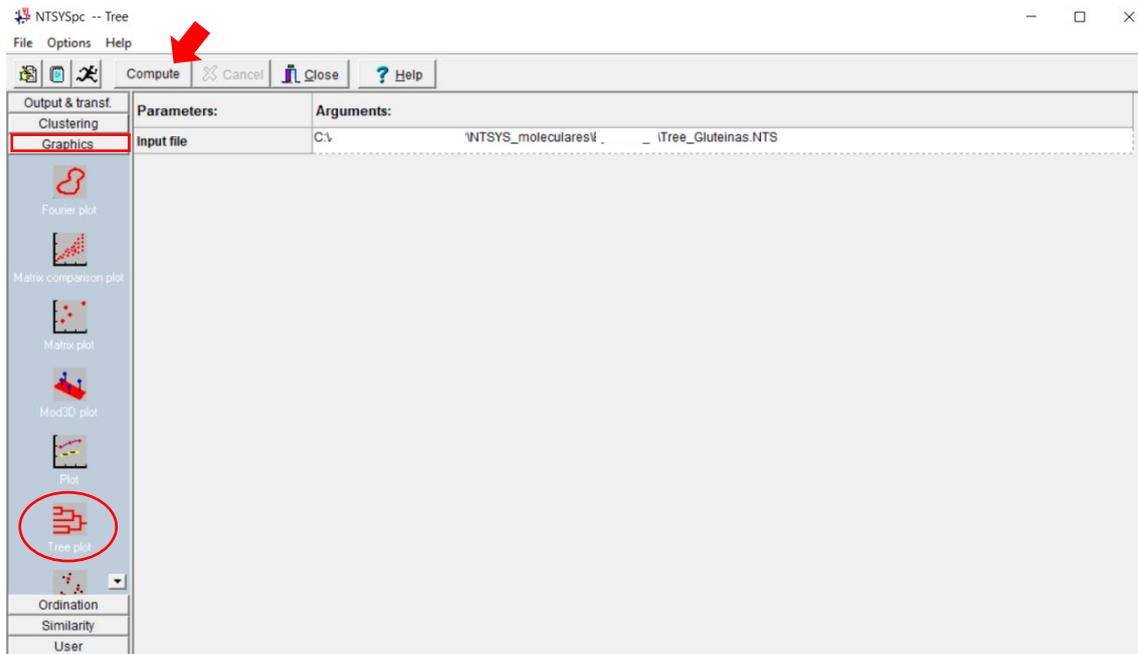
Output tree file= archivo en el que se guardarán los datos de construcción del dendrograma, Tree\_Gluteinas.nts

Clustering method= seleccionamos el método de emparejamiento que queremos emplear, en este caso utilizaremos el UPGMA que es el más empleado.

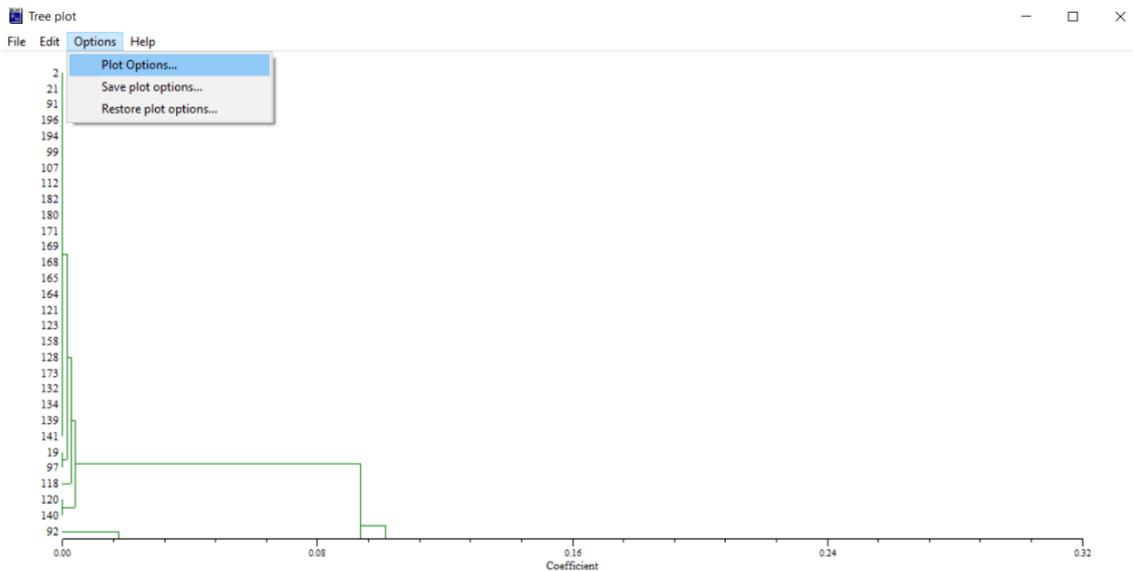
Introducidos los datos pulsaremos la tecla *Compute* y podremos visualizar el

dendrograma pulsando en la tecla de la esquina inferior izquierda 

- Además de visualizar el dendrograma a la finalización del análisis podremos visualizarlo siempre que sea necesario mediante el módulo *Graphics* → *Tree plot* y seleccionando como *Input file* el archivo que contiene los datos de agrupación, en este caso sería Tree\_Gluteinas.nts.

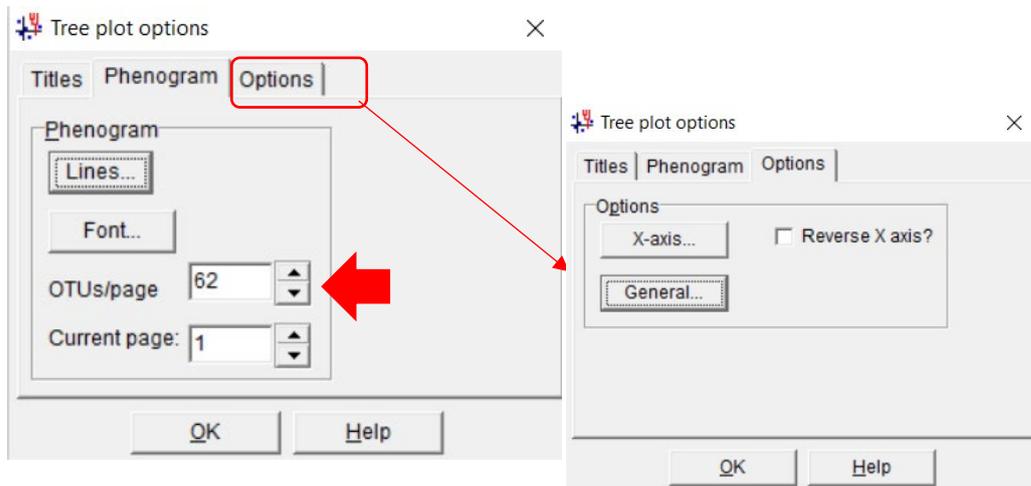


- Si el número de individuos agrupados es elevado no se podrá visualizar el dendrograma completo de forma directa, será necesario ajustar las características del mismo en el menú *Options* → *Plot Options*

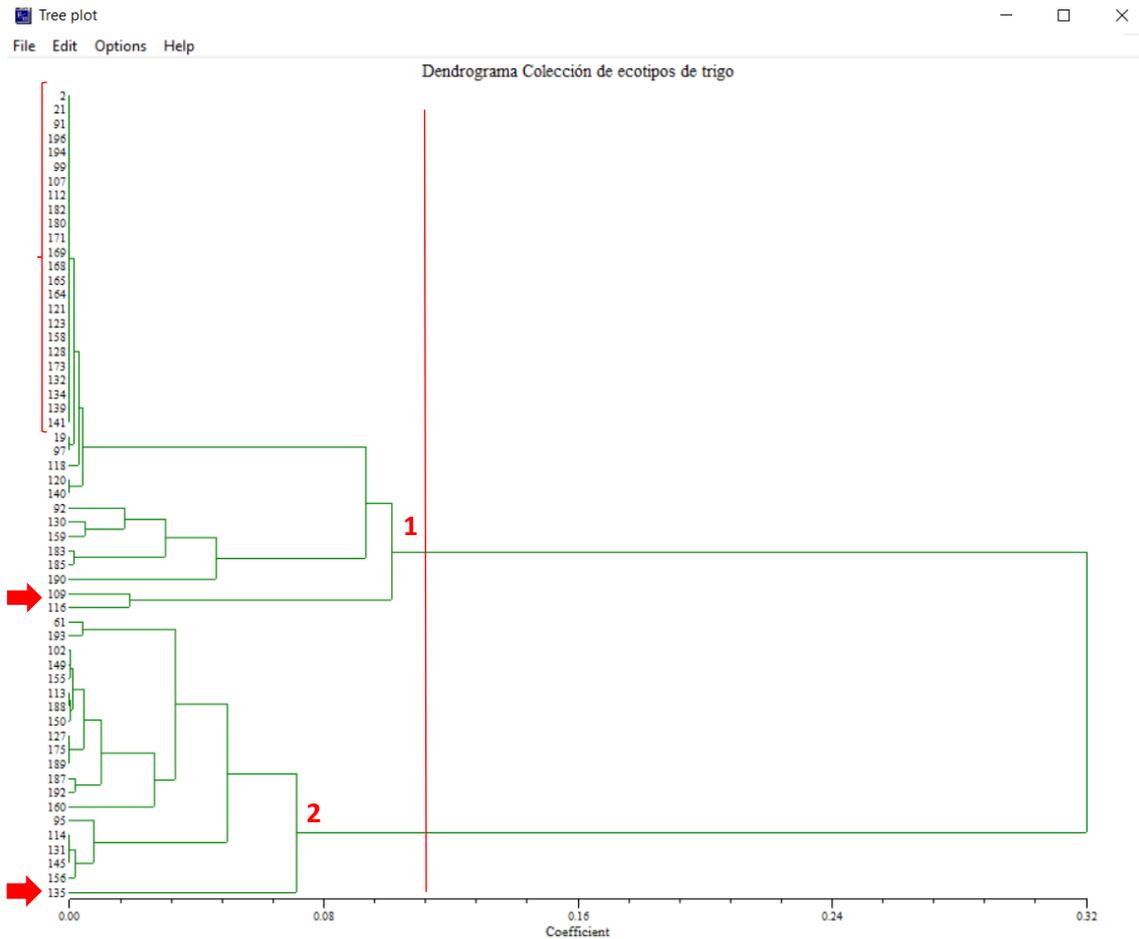


Aparecerá una nueva ventana en la que podremos indicar el número de individuos del que consta el dendrograma (en NTSYSpc los individuos objeto de análisis se los conoce como OTUs = Unidades Taxonómicamente Operativas). Además de ajustar el dendrograma para visualizarlo en su totalidad, en la ventana de opciones también se podrá insertar un título al dendrograma, modificar tipo de letra, modificar la

configuración del eje de coordenadas X en el que se representa el dendrograma, así como otras características generales del mismo.



En el menú *File* también se podrá imprimir el dendrograma obtenido o guardarlo (*Save as metafile*) en formato .EMF (Dendrograma\_ecotipos.emf), que podremos abrir mediante el programa Paint de Windows y guardarlo con el formato de imagen que se prefiera.



El dendrograma muestra la separación de la colección en dos grandes grupos, que a su vez se van subdividiendo en función de la similitud entre los ecotipos considerados. En el caso del primer grupo diferenciado, se observa un subgrupo en el que el número de ecotipos que presentan el mismo genotipo es muy numeroso. Existen genotipos como el 109, el 116 o el 135, los cuales se separan del resto de genotipos de su grupo a una mayor distancia que el resto, por lo que sus diferencias genotípicas son mayores. Este tipo de agrupación es muy útil para el estudio de las relaciones entre individuos, pero también para otros fines como, por ejemplo, la creación de *Colecciones núcleo* o *Core collections*, en las que se pretende tener la mayor variabilidad genética representada en el menor número de individuos posible, por lo que se seleccionaría uno o varios individuos de cada grupo, dependiendo de lo amplio que sea dicho grupo, para la creación de la *Core*, pudiendo obviar el empleo o la conservación del resto de individuos.

## 7. Métodos de ordenación: Análisis de Componentes Principales

La obtención de dendrogramas es el método de clasificación más adecuado cuando partimos de organismos que presentan niveles de semejanza importantes y, por tanto, es evidente que deben reunirse en grupos, de forma que nuestro problema es ver cómo se agrupan (nº de grupos, niveles de agrupamiento). Sin embargo, con frecuencia los individuos se definen mediante un conjunto de variables y es complicado determinar *a priori* si existen o no relaciones entre ellos, empleando en estos casos los métodos de ordenación.

Existen diferentes métodos que son adecuados para diferentes tipos de datos (cuantitativos o cualitativos) y que utilizan diferentes índices de distancia, pero de forma general, en todos ellos se intenta ordenar a los individuos analizados representándolos en el espacio de dimensión  $n$ , que viene dada por los datos analizados. El método de ordenación más utilizado es el Análisis de Componentes Principales (ACP) (Ringnér, 2008) con el cual se hace una transformación de los datos originales de los individuos para obtener un nuevo conjunto de variables no correlacionadas entre sí (ortogonales), llamadas Componentes Principales. Estos se calculan en un orden de importancia decreciente de tal forma que unos cuantos de los primeros componentes expliquen la mayor parte de la variación en los datos originales.

Los componentes principales contienen información en diferentes proporciones de todas las variables originales y la contribución de las variables a cada componente principal se expresa en valores y vectores propios. El valor propio representa la varianza asociada con el componente principal y decrece a medida que se generan dichos componentes. En cambio, el vector propio contiene los coeficientes de las combinaciones lineales de las  $p$  variables originales.

El ACP es una herramienta útil para analizar los datos que se generan de la caracterización y evaluación preliminar de una muestra poblacional, y permite conocer la relación existente entre las variables cuantitativas y/o cualitativas consideradas y la semejanza entre los individuos de la población, con el fin de saber cuáles variables están o no asociadas, cuáles caracterizan en el mismo sentido o en el sentido contrario y saber cómo se distribuyen los individuos según su similitud, además permite seleccionar las variables más discriminatorias para limitar el número de mediciones en caracterizaciones posteriores.

## Ejercicios. Análisis de Componentes Principales.

### Ejercicio 3.

**Fichero: Medias\_Ecotipos.xls**

#### Objetivo:

Se realizó la caracterización agronómica y de grano de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.). Se evaluaron los caracteres de grano: contenido de proteína (PROT), volumen de sedimentación del grano en SDS, peso hectolítrico (PHL) el peso de mil semillas (PMS). Agronómicamente se evaluó: la altura de planta (PALT), longitud de la espiga (ELON), número de espiguillas por espiga (EESPN), densidad de la espiga (EDEN), y días transcurridos desde la siembra hasta la fecha de espigado (DESP) y de madurez fisiológica (DMAD) y rendimiento (RDTO).

#### Cuestiones:

1. Determinar cuáles de los caracteres empleados en la caracterización son los más discriminantes para la caracterización de la colección.
2. Estudiar la estructura de la colección para determinar la existencia de ecotipos de características similares, ver las relaciones entre los mismos y descartar la existencia de duplicados.

#### Resolución Ejercicio 3.

- Introducción de la matriz de datos base con las medias por carácter de los ecotipos evaluados (Medias\_Ecotipos.xls) en el NTSYSpc a través del editor NTedit para guardar los archivos en formato “.nts” (ver apartado 4.2 *Introducción de la matriz de datos en el NTSYSpc*).
- Estandarización de la matriz de datos.

Generalmente las variables seleccionadas para la caracterización tienen diferentes unidades de medida, por ello, en el análisis multivariante es común estandarizar las variables para garantizar una comparabilidad precisa y equilibrada entre ellas.

La estandarización se realizará a través del módulo *Output & transf.* dentro de este módulo podemos emplear las opciones:

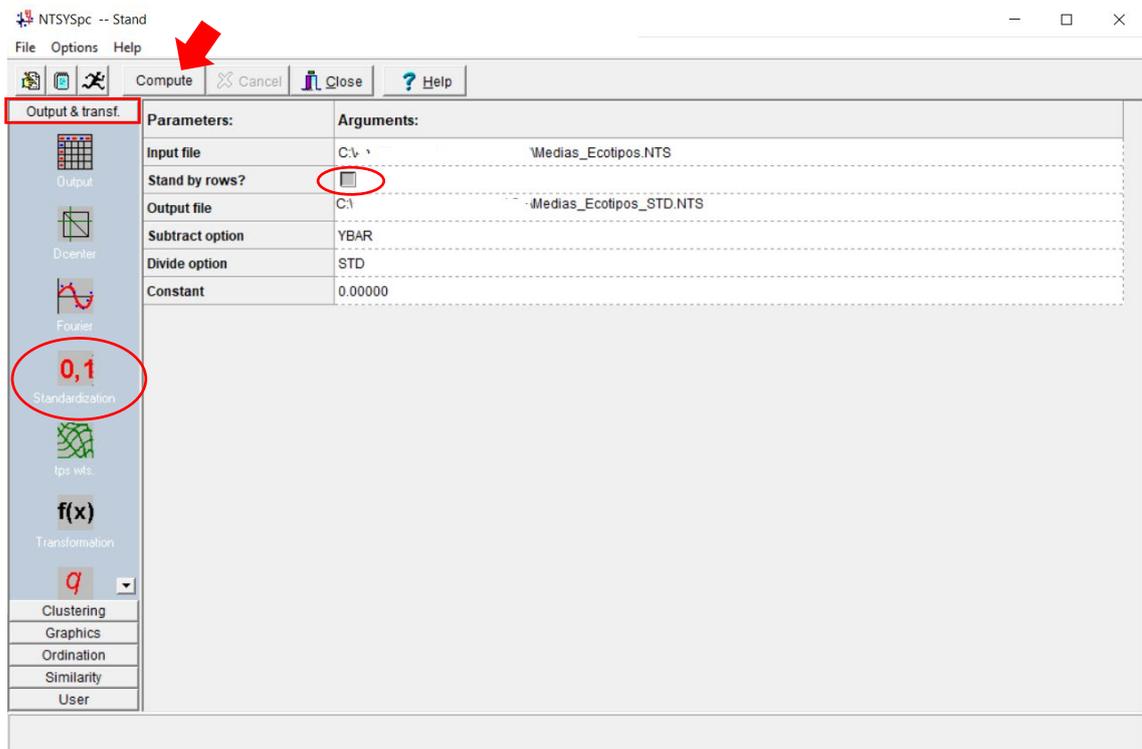
→ *Standardization*, que realiza transformaciones lineales del tipo:

$$y = \frac{y - \text{substraction option}}{\text{divide option}}$$

→ *Transformation*, que realiza transformaciones logarítmicas empleando el código LOG1

En este ejercicio utilizaremos el módulo de *Standardization*.

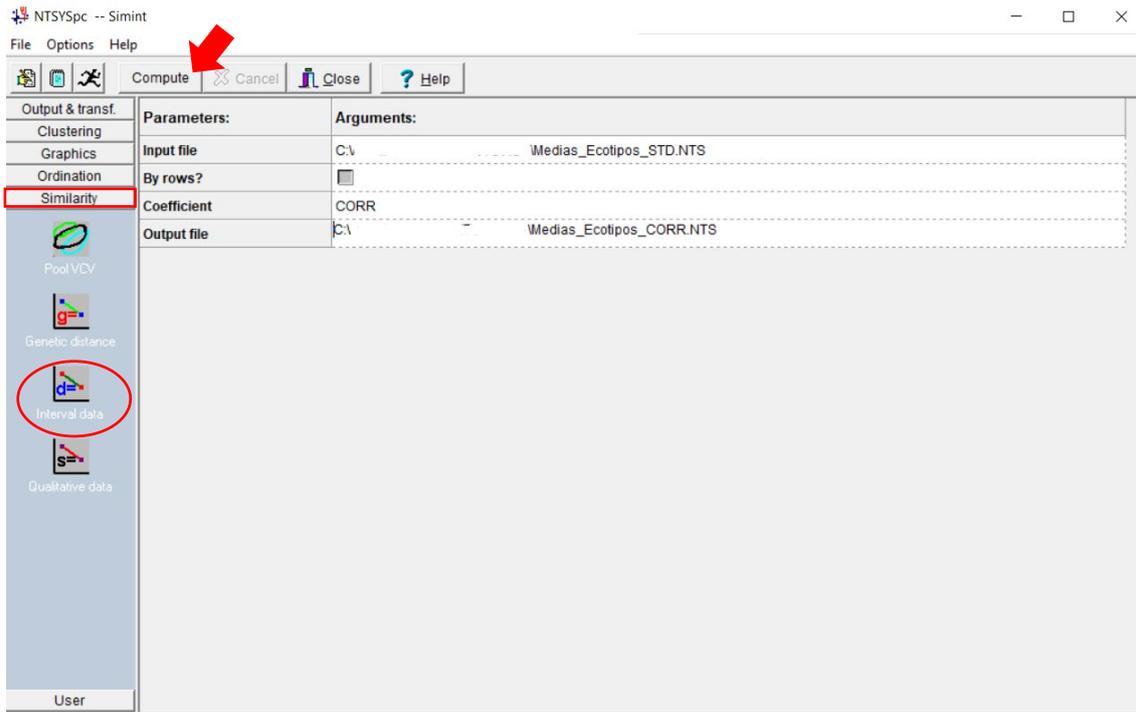
Como las variables en la matriz de datos se encuentran en columna debemos dejar sin marcar la casilla de *Stand by rows?*



- Calcular la matriz de correlaciones entre caracteres

A partir de la matriz de datos base estandarizada (Medias\_Ecotipos\_STD.nts) se calcula la matriz de correlación, que es un tipo de matriz de similitud (Medias\_Ecotipos\_CORR.nts). Para ello empleamos el módulo *Similarity* → *Interval data*, *Coefficient CORR*.

[También se podría calcular la matriz de varianzas-covarianzas a partir de la matriz de datos estandarizada (Medias\_Ecotipos\_STD.nts), en la que se dan mayores pesos a las variables que presentan varianzas mayores. Para su cálculo empleamos el módulo *Similarity* → *Interval data*, *Coefficient VARCOV*]

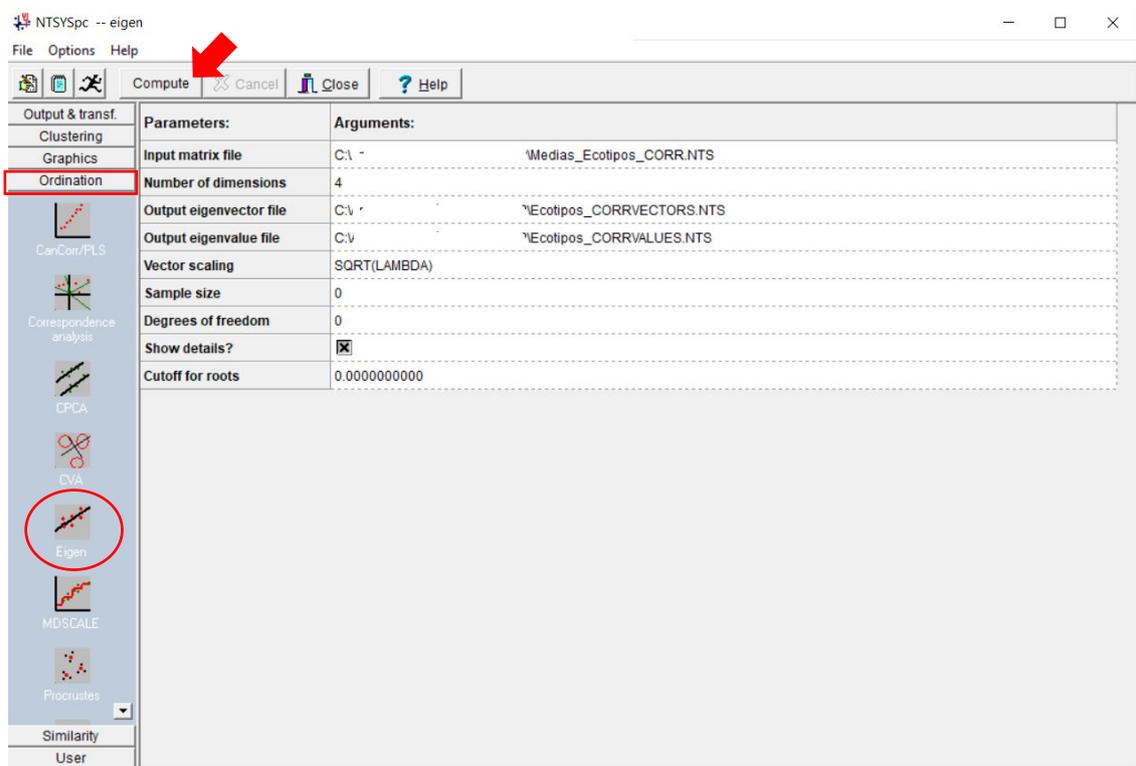


- Obtener los componentes principales

A partir de la matriz de correlaciones calculamos los vectores

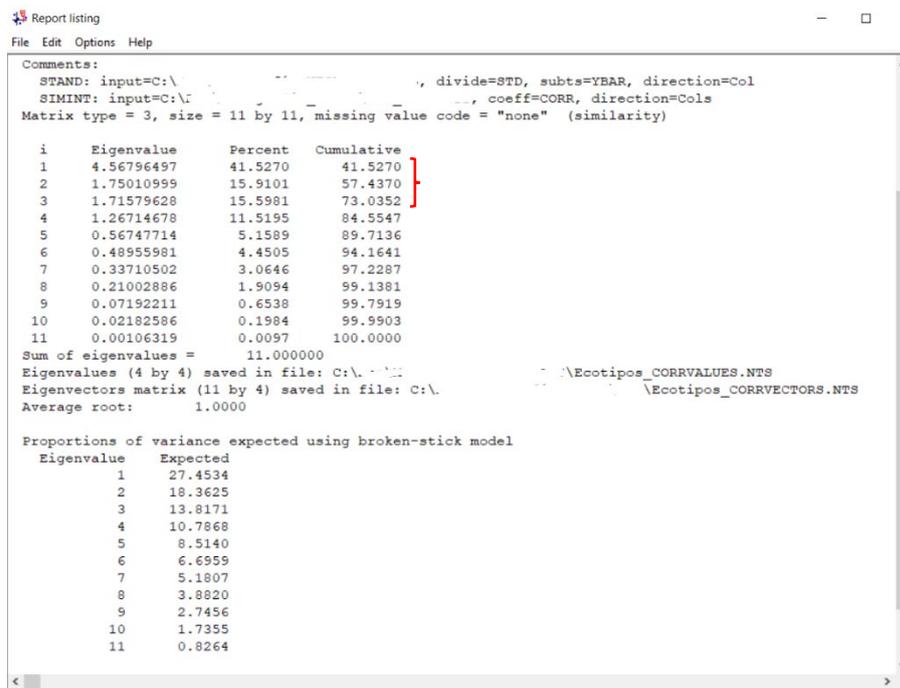
(Ecotipos\_CORRVECTORS.nts) y los valores propios (Ecotipos\_CORRVALUES.nts)

mediante el módulo *Ordination* → *Eigen*



En la ventana del *Reporting list* se podrán visualizar los valores propios correspondientes a cada uno de los componentes principales y los porcentajes de variación que recogen. El % de explicación se obtiene dividiendo el valor propio de cada componente principal por la suma de los valores propios y se dan en orden descendente, de forma que el primer eje es el que reúne mayor cantidad de información, seguido del segundo, y así sucesivamente.

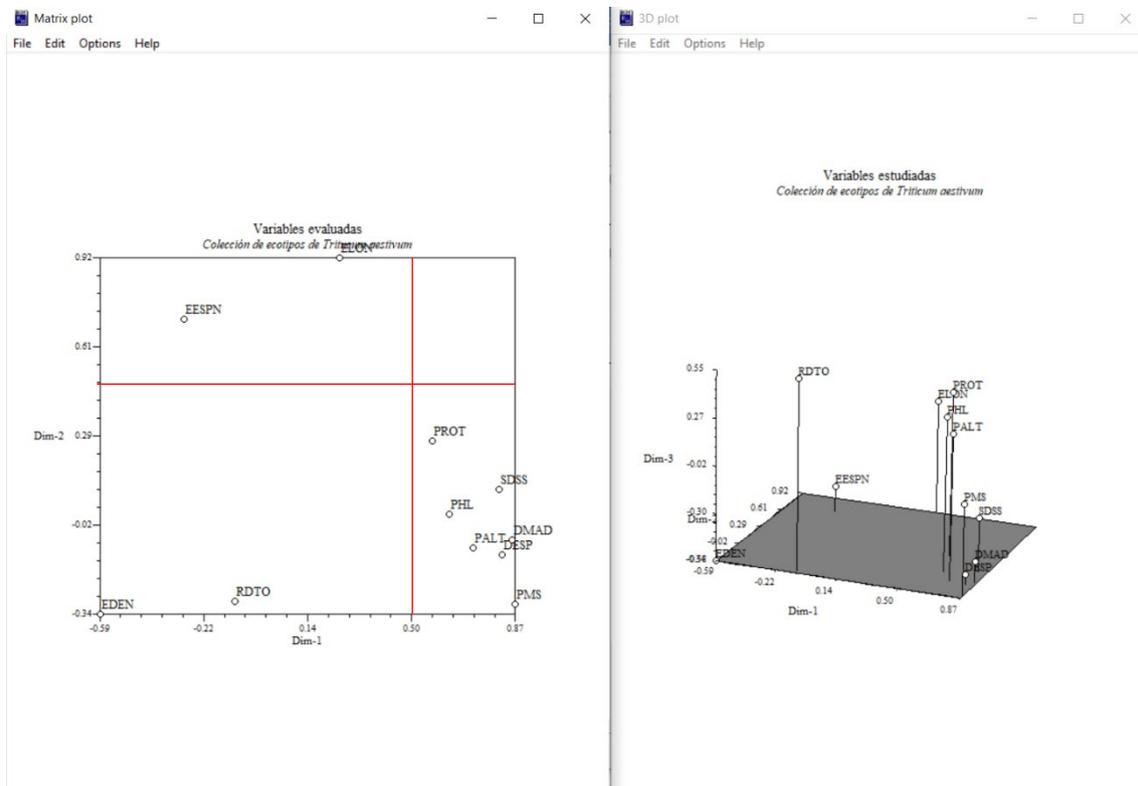
En este caso en particular prácticamente el 42% de la variabilidad encontrada en la matriz de datos base se explica mediante la primera componente principal y considerando las tres primeras componentes tendríamos la explicación al 73% de la variabilidad.



En la esquina inferior izquierda podremos ver los iconos  en los que se visualizarán los gráficos de dispersión bidimensionales o tridimensionales del espacio factorial, donde se puede observar de manera gráfica la contribución de cada carácter o variable a cada uno de los componentes principales. Las variables que están próximas entre sí tendrán un nivel de varianza similar y las que alcancen los mayores valores en el eje serán las que más contribuyen a la variabilidad explicada por la componente principal.

En este caso, los caracteres de madurez fisiológica (DESP, DMAD) y los caracteres de semillas PMS y SDSS tienen un peso importante en la variabilidad explicada por el primer componente principal, siendo el carácter de la espiga EDEN de peso, pero en el sentido contrario que los caracteres anteriores. Considerando el segundo componente principal, el carácter de longitud de la espiga (ELON) es el que presenta mayor peso a la hora de explicar la variabilidad y el rendimiento aportaría peso en el sentido contrario, es decir, a mayor peso, menor rendimiento y viceversa.

La distribución de los caracteres indica que todos los estudiados tienen importancia en la determinación de la variabilidad, estando todos por encima de 0,5 tanto en la primera como en la segunda componente principal.

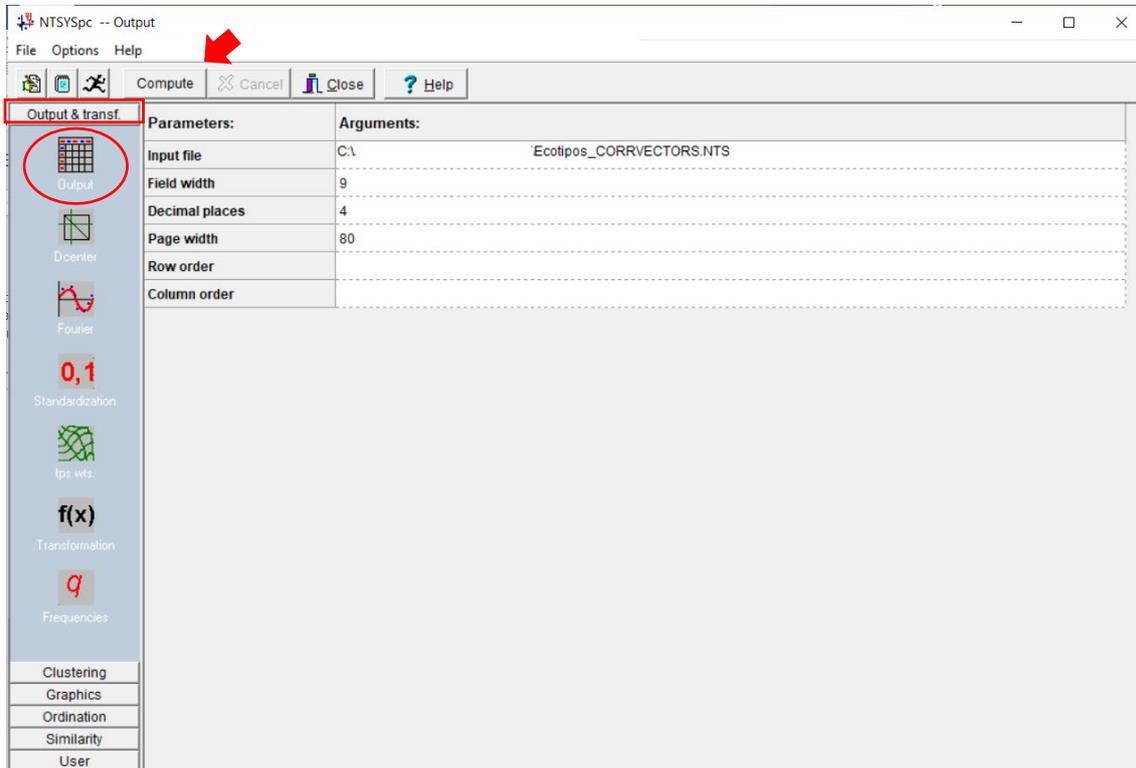


En el menú *Options* → *Plot options* se podrán modificar las características de visualización de los gráficos según nos interese.

En el menú *File* → *Save as metafile* podremos guardar los gráficos en formato .EMF, (Ficheros Plot 2D\_Variables.emf y Plot 3D\_Variables.emf) que podremos abrir mediante el programa Paint de Windows y guardarlos con el formato de imagen que se prefiera.

- Obtener los eigen-vectors

Debemos solicitar al programa que nos muestre los valores de los vectores calculados representando el archivo Ecotipos\_CORRVECTORS.nts mediante el módulo *Output & transf.* → *Output*.



En la ventana del *Reporting list* podremos visualizar los eigen-vectors con los que se representa cada variable.

```

Report listing
File Edit Options Help
Output: NTSYSpc 2.11a, (C) 2000-2002, Applied Biostatistics Inc.Date & time: 10/10/2002 17:52:50
-----
Input parameters
Read input from file: C:\... \Ecotipos_CORRVECTORS.NTS
Format: width=9 decimals=4
Page width: 80
Field width: 9
Decimal places: 4
Page width: 80
Comments:
STAND: input=C:\... , divide=STD, subts=YBAR, direction=Col
SIMINT: input=C:\... , coeff=CORR, direction=Cols
EIGEN: input=C:\... \Medias_Ecotipos_CORR.NTS, k=4 vectors, length=SQRT(LAMBDA)
Matrix type = 1, size = 11 by 4, missing value code = "none" (rectangular)

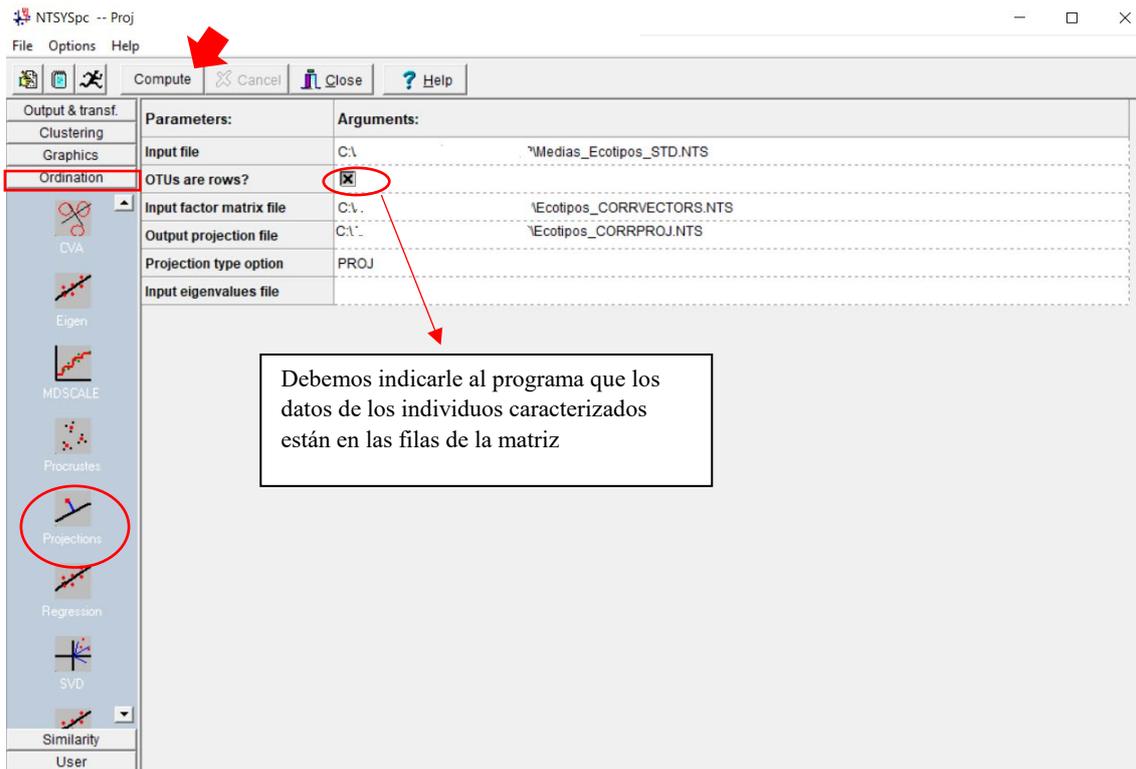
      C1      C2      C3      C4
-----
PHL | 0.6367  0.0154  0.3337  0.4320
PROT | 0.5749  0.2741  0.3909 -0.4881
SDSS | 0.8089  0.1012 -0.2621 -0.0557
PMS | 0.8669 -0.3014 -0.0429  0.2028
PALT | 0.7182 -0.1043  0.2802  0.0403
ELON | 0.2519  0.9206  0.1020  0.1982
EESPN | -0.2935  0.7041 -0.4233  0.4383
EDEN | -0.5880 -0.3351 -0.5785  0.2100
DESP | 0.8191 -0.1278 -0.5196  0.0533
DMAD | 0.8548 -0.0755 -0.4563  0.1018
RDTO | -0.1160 -0.2921  0.5479  0.7126

```

Como se observaba en el grafico 2D, los caracteres de mayor peso en la primera componente principal serían los de madurez fisiológica y de semilla PMS y SDSS, por tanto, sería un componente de madurez y peso de semilla. En el caso de la segunda componente principal, destaca el carácter de longitud de la espiga (ELON), seguido del número de espiguillas por espiga (EESPN), por tanto, sería un componente de característica de la espiga. En el caso de la tercera componente principal, sería un componente de rendimiento, carácter que presenta el valor más cercano a 1.

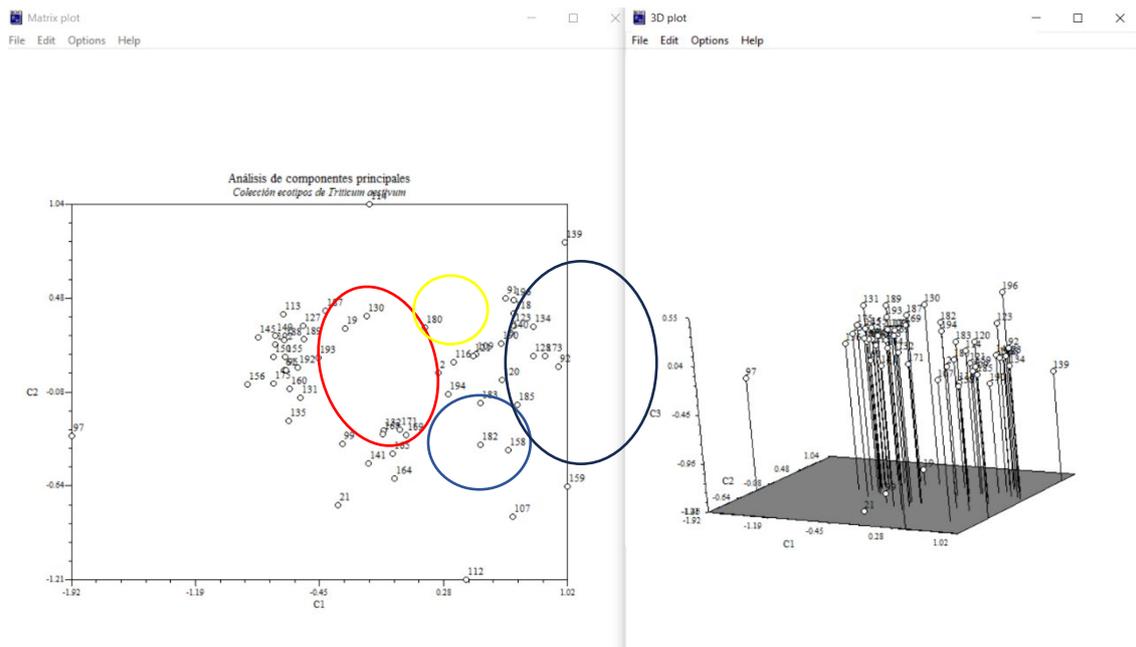
- Proyección de los individuos sobre los vectores calculados

Conocida la relación de los caracteres originales con los componentes principales calculados, se realiza la proyección tridimensional de los individuos caracterizados sobre el conjunto de ejes definidos por los componentes principales debemos utilizar el módulo *Ordination* → *Projections* para realizar la proyección de la matriz de datos estandarizada (Medias\_Ecotipos\_STD.nts).



En la esquina inferior izquierda podremos ver los iconos  en los que se obtienen los diagramas de dispersión bidimensionales o tridimensionales de proyección de los individuos caracterizados sobre los componentes principales. En el menú *Options* → *Plot options* se podrán modificar las características de visualización de los gráficos según nos interese.

En el menú *File* → *Save as metafile* podremos guardar los gráficos en formato .EMF, (Ficheros: Plot 2D\_Ecotipos.emf y Plot 3D\_Ecotipos.emf) que podremos abrir mediante el programa Paint de Windows y guardarlos con el formato de imagen que se prefiera.



En el gráfico representativo de las dos primeras componentes principales, se observa *grosso modo* el agrupamiento de los ecotipos en 4 grandes grupos y ecotipos claramente diferentes al resto. El grupo de la izquierda, marcado en rojo, es un grupo más compacto, mientras que el grupo más a la derecha, marcado en azul oscuro, es más disperso, pudiendo hacer nuevas agrupaciones dentro del mismo.

#### Ejercicio 4. Análisis de componentes principales. Distancia de Gower

Fichero: DIST\_GOWER.xls

##### Objetivo:

Se realizó la caracterización de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.) mediante la evaluación agronómica y de grano de caracteres tanto cuantitativos como cualitativos. Dado el carácter mixto de las variables consideradas se calculó la matriz de distancia empleando la distancia de Gower.

##### Cuestiones:

Dada la matriz de distancias entre ecotipos realizar la agrupación de los mismo mediante un análisis de componentes principales.

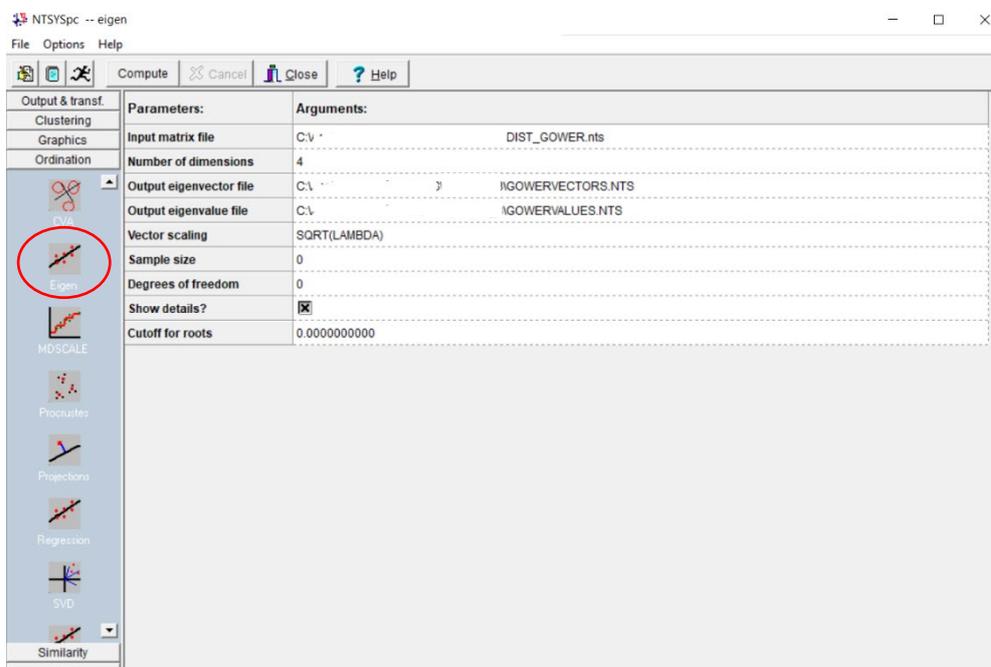
##### Resolución Ejercicio 4.

- Introducción de la matriz de distancias (DIST\_GOWER.xls) en el NTSYSpc a través del editor NTedit para guardar los archivos en formato “.nts” (ver apartado 4.2 *Introducción de la matriz de datos en el NTSYSpc*). Al no tratarse de una matriz de datos rectangular sino de una matriz de distancias debemos indicarlo en la cabecera.

	A	B	C	D	E	F	G	H	I	J
1	3	57	57	0						
2	2	19	21	61	91	92	95	97	99	
3	2	1								
4	19	0.39871	1							
5	21	0.41623	0.47111	1						
6	61	0.46968	0.40216	0.37012	1					
7	91	0.40766	0.38356	0.34515	0.41697	1				
8	92	0.42587	0.3711	0.32593	0.40139	0.5078	1			
9	95	0.4241	0.38087	0.33917	0.55698	0.38293	0.47967	1		

- Obtención de los eigen-vectores

En ocasiones no partimos de una matriz de datos base rectangular, sino que debemos partir de una matriz con datos de similitud o de distancias entre los individuos evaluados porque la hemos calculado mediante un software distinto al NTSYSpc. En dichos casos, el análisis de componentes principales se simplifica y se pasaría directamente al paso de obtención de los valores propios (GOWERVALUES.nts) y vectores (GOWERVECTORS.nts) mediante el módulo *Ordination* → *Eigen*

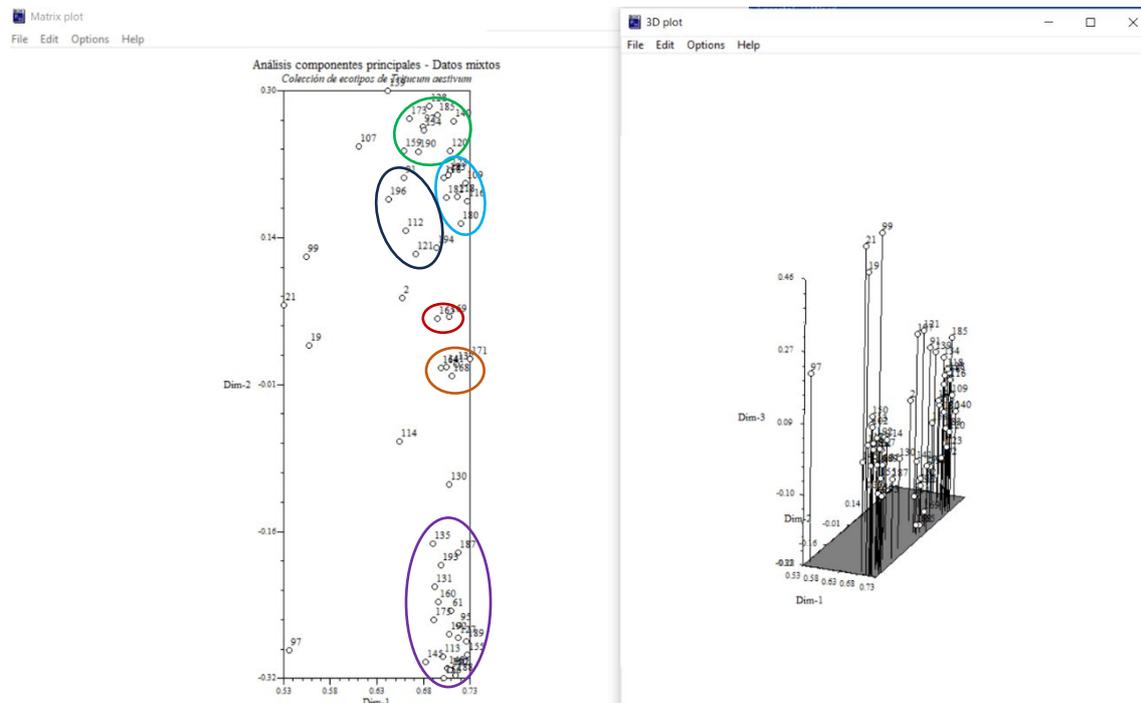




En la esquina inferior izquierda podremos ver los iconos en los que se obtienen los diagramas de dispersión bidimensionales o tridimensionales de proyección de los individuos caracterizados sobre los componentes principales.

En el menú *Options* → *Plot options* se podrán modificar las características de visualización de los gráficos según nos interese.

En el menú *File* → *Save as metafile* podremos guardar los gráficos en formato .EMF, (Ficheros: Plot 2D\_Ecotipos\_DIST.emf y Plot 3D\_Ecotipos\_DIST.emf) que podremos abrir mediante el programa Paint de Windows y guardarlos con el formato de imagen que se prefiera.



Como se observa en el gráfico, se obtienen varios grupos de ecotipos definidos, así como ecotipos que aparecen claramente separados de las agrupaciones. Si se compara este gráfico con el gráfico del ACP del ejercicio 2, se observa que la definición de los grupos es más clara en este caso, por lo que, la utilización de caracteres de distinto tipo (cualitativos + cuantitativos) es muy útil para realizar caracterizaciones más precisas.

Podemos solicitar al programa que nos muestre los valores de los vectores calculados representando el archivo GOWERVECTORS.nts mediante el módulo *Output & transf.* → *Output*, obteniendo en la ventana de *Reporting list* los vectores que representan a cada individuo analizado en cada uno de los componentes principales.

Los datos de los vectores visualizados podrían copiarse y pegarse en una hoja de Excel para realizar la representación de los datos mediante un gráfico de dispersión XY, esto nos permite hacer una presentación más vistosa de los resultados.

```

Report listing
File Edit Options Help
Output: NTSYSpc 2.11a, (C) 2000-2002, Applied Biostatistics Inc.Date & time:
-----
Input parameters
Read input from file: C:\          \GOWERVECTORS.NTS
Format: width=9 decimals=4
Page width: 80
Field width: 9
Decimal places: 4
Page width: 80
Comments:
EIGEN: input=C:\          \DIST_GOWER.nts, k=4 vectors, length=SQRT(LAMBDA)
Matrix type = 1, size = 57 by 4, missing value code = "none" (rectangular)

      C1      C2      C3      C4
-----
2 | 0.6585  0.0815  0.0407  0.1393
19 | 0.5614  0.0312  0.3821  0.2923
21 | 0.5349  0.0736  0.4360  0.4083
61 | 0.7097 -0.2487  0.0224  0.0465
91 | 0.6608  0.2079  0.1434 -0.1232
92 | 0.6800  0.2617  0.0317 -0.1047
95 | 0.7186 -0.2637 -0.0925 -0.0136
97 | 0.5404 -0.2896  0.2092  0.0920
99 | 0.5589  0.1246  0.4606  0.2611
102 | 0.7105 -0.3106  0.1020 -0.0934
107 | 0.6131  0.2413  0.1640  0.1732
109 | 0.7245  0.2027  0.0267 -0.0545
112 | 0.6621  0.1519 -0.2122  0.0980
113 | 0.7012 -0.2962  0.1025 -0.1077
114 | 0.6557 -0.0698 -0.0181 -0.1289
116 | 0.7264  0.1838  0.0733 -0.1131
118 | 0.7164  0.1881  0.0997 -0.2454
120 | 0.7084  0.2367 -0.0873 -0.0827
121 | 0.6727  0.1278  0.2159 -0.1078
123 | 0.7091  0.2160 -0.1221 -0.0338
127 | 0.7175 -0.2764  0.0347 -0.1451
128 | 0.6871  0.2839  0.0567 -0.1488
130 | 0.7074 -0.1153 -0.0452 -0.1107
131 | 0.6928 -0.2232 -0.1044 -0.0192
132 | 0.7155  0.0115 -0.1396  0.1709

```

## 8. Distancias geográficas

Al definir la variabilidad genética de las poblaciones se introdujo el concepto de *Flujo genético* como una de las principales fuentes generadoras de variabilidad. Las migraciones de parte de una población o de poblaciones completas y el mestizaje con las poblaciones de las zonas receptoras genera la aparición de nuevos genotipos que evolucionan diferenciándose en nuevas poblaciones, en cuya diferenciación influye enormemente el medioambiente en el que dichas poblaciones evolucionan. Por ello, es importante tener en cuenta la distribución geográfica y la influencia que la misma pudiese tener en las diferencias genéticas y morfológica que observamos en los individuos de las poblaciones a estudio.

Para determinar si la distancia genética en poblaciones o individuos dentro de poblaciones está relacionada con la distancia geográfica entre los mismos se emplea el Test de Mantel.

### **9. Test de Mantel**

El test de Mantel (Mantel, 1967) es un test estadístico aplicado para la estimación del grado de correlación existente entre dos matrices, X e Y, donde ambas matrices deben ser del mismo rango.

La hipótesis nula postula que las distancias/similitudes entre las variables de la matriz respuesta Y no están linealmente correlacionadas con las correspondientes distancias/similitudes en la matriz modelo X. Se trata, por tanto, de evaluar si la asociación (positiva o negativa) es más robusta de lo que cabría esperar por puro azar. El estadístico del test de Mantel Z se calcula mediante la suma de los productos cruzados de los valores de las dos matrices de similitud/distancia, excluyendo la diagonal principal:

$$Z = \sum X_{ij} Y_{ik}$$

donde  $X_{ij}$  e  $Y_{ik}$  son los elementos no diagonales de las matrices X (o S= semejanza/distancias genéticas, etc.) e Y (o C= valores cofenéticos/ distancias genéticas, etc.)

Normalmente, se emplea el coeficiente estandarizado de Mantel (r), que se calcula mediante el producto cruzado de las dos matrices dividido por  $[(n(n-1)/2)-1]$ , donde n son los objetos comparados. El coeficiente r de Mantel se puede interpretar como un coeficiente de correlación y su valor oscila entre -1 y +1.

Dadas las características de este test, es muy utilizado en el estudio de las relaciones entre las especies y variables ambientales. Aunque también se puede emplear, entre otros, para validar los resultados obtenidos en los métodos de agrupación, cuantificando la distorsión debida al método de agrupación, o para comparar la similitud entre dos métodos de caracterización.

## Ejercicios. Test de Mantel.

### Ejercicio 5.

**Ficheros:** Coordenadas\_Ecotipos.xls, Medias\_Ecotipo\_STD.nts

#### Objetivo:

Se dispone de datos de la caracterización de una colección de 57 ecotipos de trigo (*Triticum aestivum*) mediante evaluación morfoagronómica, así como los datos gps (latitud, longitud y altura) de localización de los puntos de recolección de cada ecotipo.

#### Cuestión:

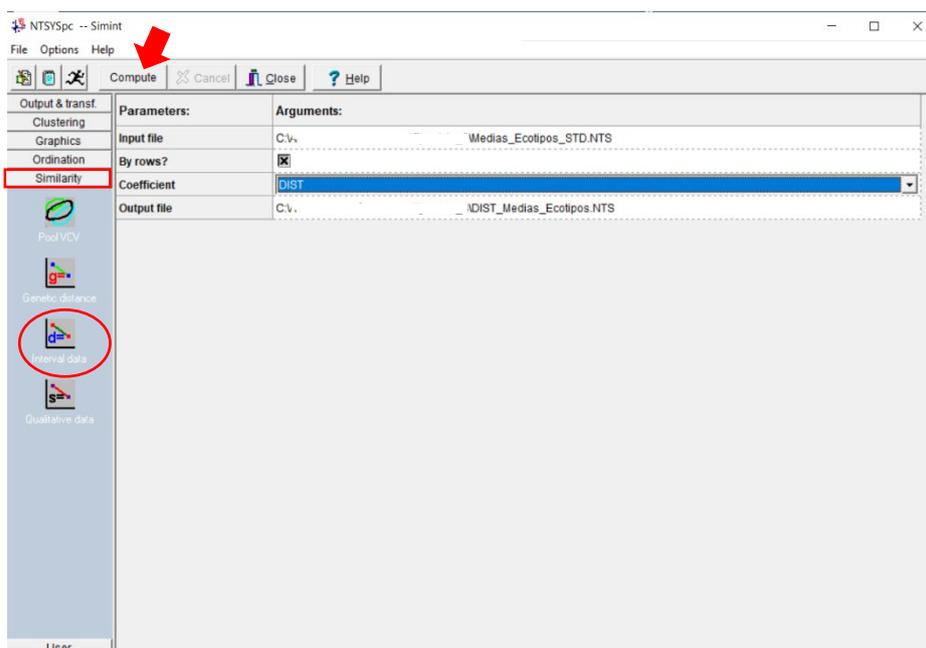
Determinar si existe un patrón geográfico que explique las diferencias entre los individuos de la colección.

### Resolución Ejercicio 5.

- Introducir la matriz con las coordenadas geográficas de los ecotipos (Coordenadas\_Ecotipos.xls) en el NTSYSpc a través del editor NTedit para guardar los archivos en formato “.nts” (ver apartado 4.2 *Introducción de la matriz de datos en el NTSYSpc*).

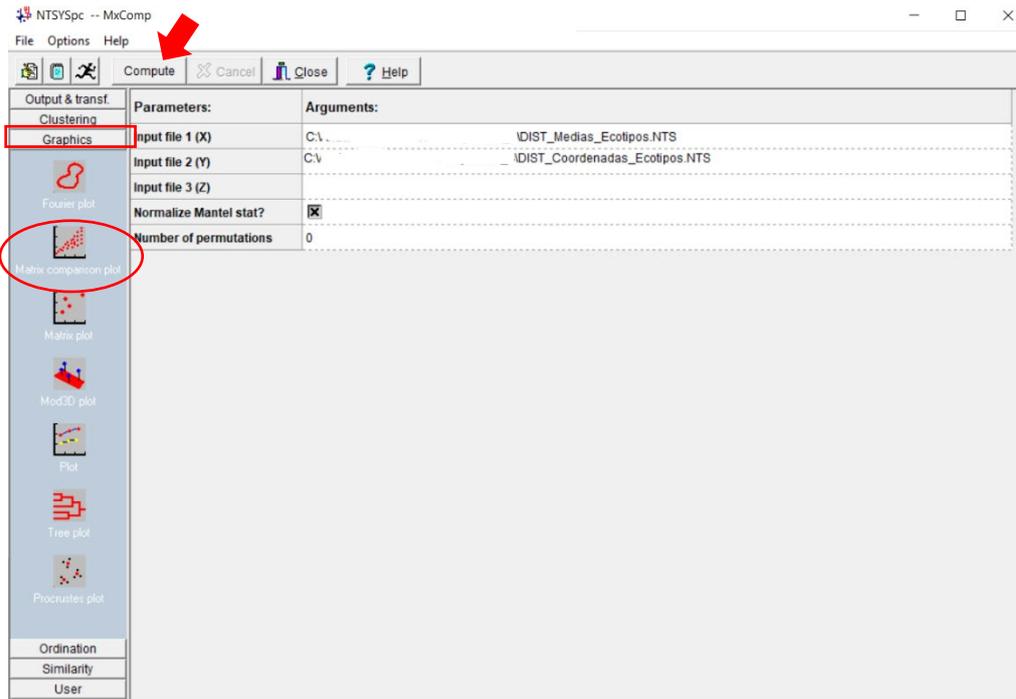
- Obtención de la matriz de distancias

A partir de la matriz de datos estandarizada obtenida en el ejercicio 3 (Medias\_Ecotipo\_STD.nts) calcularemos la matriz de distancias (DIST\_Medias\_Ecotipos.nts) mediante el módulo *Similarity* → *Interval data*, *Coefficient DIST*

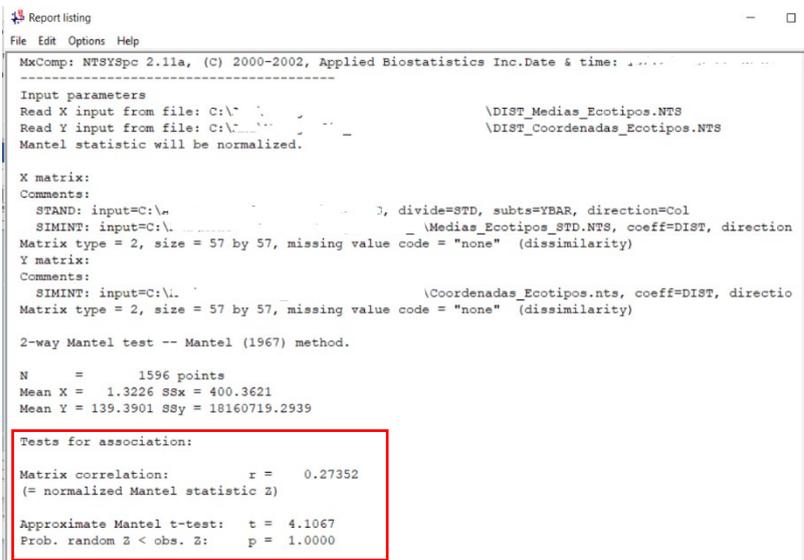


A partir de la matriz con los datos geográficos (Coordenadas\_Ecotipos.nts) calcular la matriz de distancias (DIST\_Coordenadas\_Ecotipos.nts) por el procedimiento seguido con la matriz de medias de los ecotipos.

- Realizar la comparación entre matrices
- Mediante el módulo *Graphics* → *Matrix comparison plot* realizaremos el test de Mantel para comparar las matrices de distancia calculadas.



En la ventana de *Reporting list* se visualizarán los resultados de la comparación de matrices



El coeficiente  $r$  de Mantel se interpretaría como un coeficiente de correlación, de forma que valores de 0 implicaría la ausencia de correlación y los valores cuanto más próximos a 1 indicarían la existencia de una gran correlación. Así, los valores próximos a 1 indicaría una correlación entre los datos morfoagronómicos y los geográficos, indicando la existencia de un patrón geográfico.

En este caso, el valor del coeficiente de Mantel  $r$  tiene un valor de 0,27, por lo cual, la variabilidad encontrada en la colección de ecotipos no se depende de su lugar de recolección ya que no se observa un patrón de distribución geográfica.

### **Ejercicio 6.**

**Fichero:** GD\_Gluteinas.nts, Tree\_Gluteinas.nts

#### **Objetivo:**

Se realizó el análisis clúster de una colección de 57 ecotipos de trigo (*Triticum aestivum* L.) empleando datos moleculares.

#### **Cuestión:**

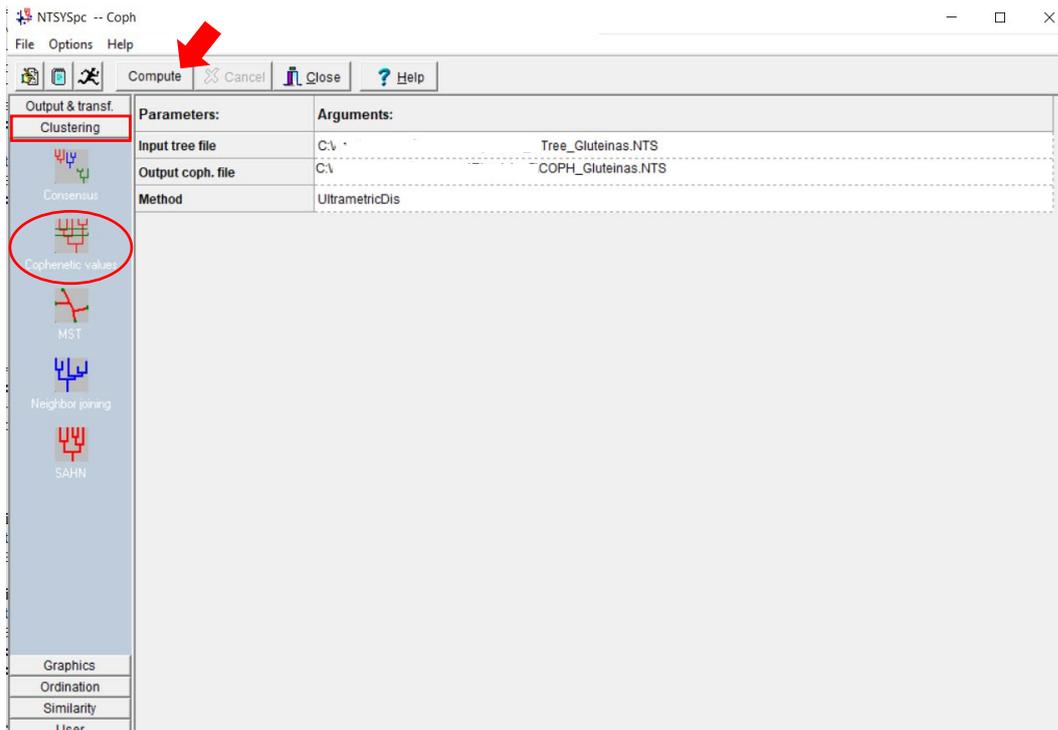
Determinar si la agrupación obtenida en el análisis clúster representa las similitudes entre los ecotipos.

### **Resolución Ejercicio 6.**

- Obtención de la matriz cofenética

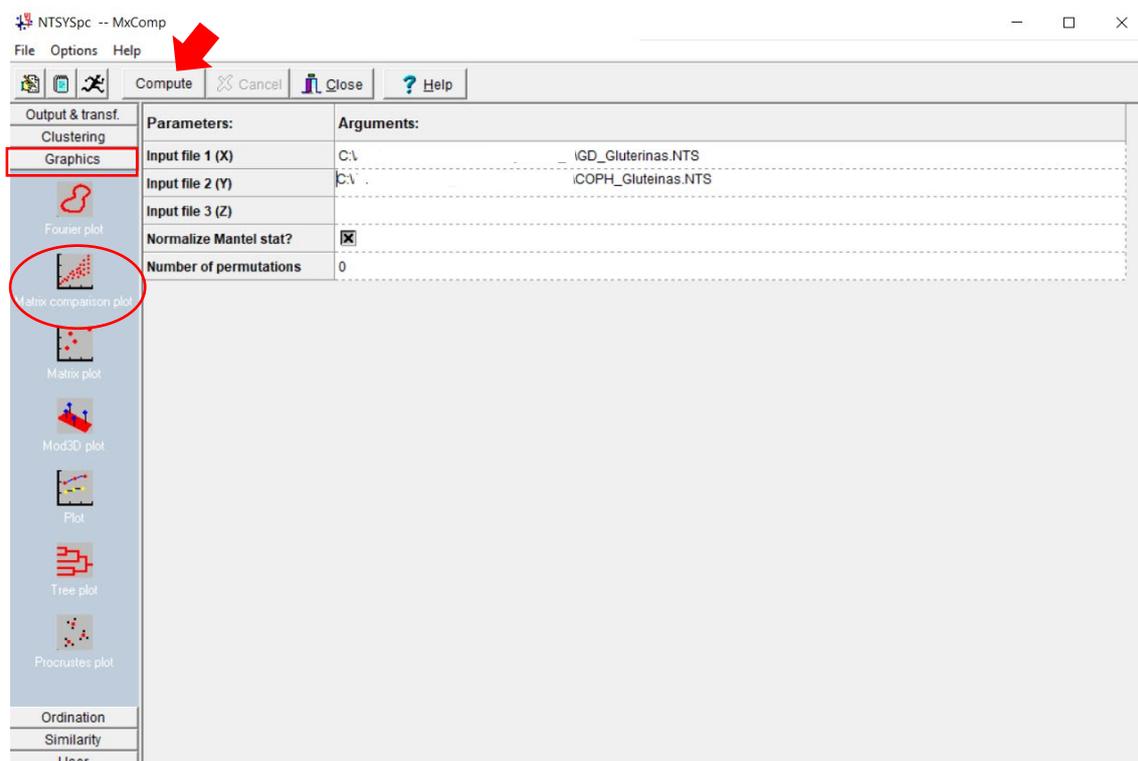
Para determinar que la estructura obtenida en el análisis clúster es representativa de las similitudes entre los individuos analizado se emplea el coeficiente de correlación cofenética (Sokal y Rohlf, 1962). Este coeficiente mide la correlación entre las distancias obtenidas a partir de los datos originales y las distancias finales obtenidas tras el proceso de agrupación iniciales y las obtenidas

Para el cálculo de la matriz cofenética utilizamos el módulo *Clustering* → *Cophenetic values* utilizando la matriz de datos obtenida en el ejercicio 2, Tree\_Gluteinas.nts

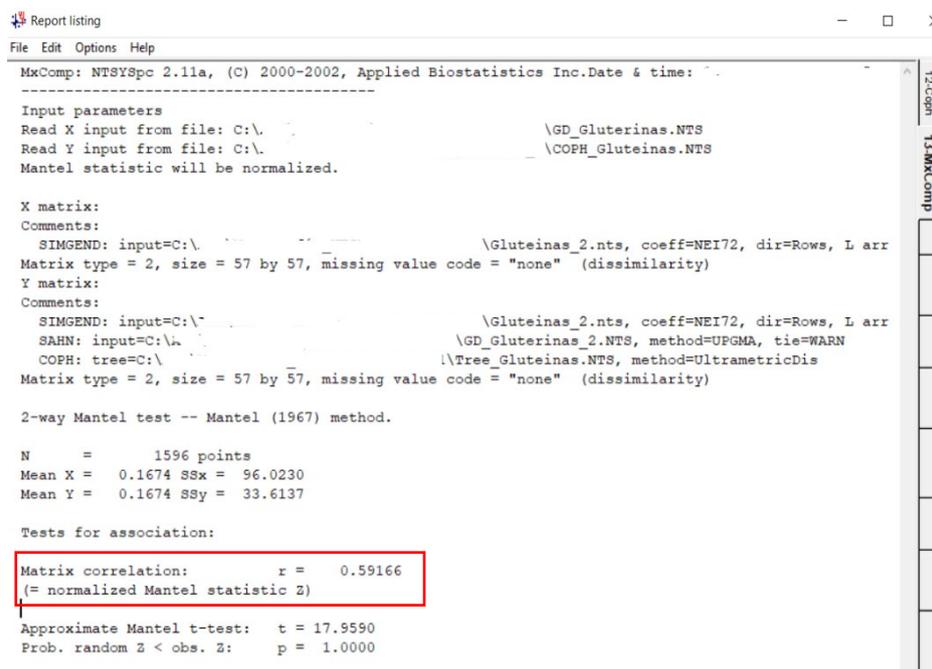


- Comparación de matrices

Mediante el módulo *Graphics* → *Matrix comparison plot* realizaremos el test de Mantel para comparar la matriz de datos cofenéticos (COPH\_Gluteinas.nts) y la matriz de distancias genéticas (obtenida en el ejercicio 2, GD\_Gluteinas.nts).



En la ventana de *Reporting list* se visualizarán los resultados de la comparación de matrices.



```
Report listing
File Edit Options Help
MxComp: NTSYSpc 2.11a, (C) 2000-2002, Applied Biostatistics Inc. Date & time: ...
-----
Input parameters
Read X input from file: C:\... \GD_Gluterinas.NTS
Read Y input from file: C:\... \COPH_Gluteinas.NTS
Mantel statistic will be normalized.

X matrix:
Comments:
  SIMGEND: input=C:\... \Gluteinas_2.nts, coeff=NEI72, dir=Rows, L arr
Matrix type = 2, size = 57 by 57, missing value code = "none" (dissimilarity)
Y matrix:
Comments:
  SAHN: input=C:\... \GD_Gluterinas_2.NTS, method=UPGMA, tie=WARN
  COPH: tree=C:\... \Tree_Gluteinas.NTS, method=UltrametricDis
Matrix type = 2, size = 57 by 57, missing value code = "none" (dissimilarity)

2-way Mantel test -- Mantel (1967) method.

N      =      1596 points
Mean X = 0.1674 SSx = 96.0230
Mean Y = 0.1674 SSy = 33.6137

Tests for association:
Matrix correlation:      r = 0.59166
(= normalized Mantel statistic Z)

Approximate Mantel t-test:  t = 17.9590
Prob. random Z < obs. Z:  p = 1.0000
```

Los valores más próximos a 1 indicarían un ajuste muy bueno entre la correlación cofenética y las distancias genéticas y, por tanto, un agrupamiento clúster que se ajusta a la estructura dada por la matriz de datos.

El coeficiente  $r$  en nuestro análisis es de 0,59, lo que indicaría que el ajuste es pobre, por lo que el agrupamiento realizado no nos está representando bien las similitudes entre individuos. En estos casos sería necesario realizar el análisis clúster con más información por individuo, así que se deberían analizar más loci.

## 10. Bibliografía

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation Procedures. *The American Journal of Human Genetics*, 19: 233-257.

Cavalli-Sforza, L.L., Bodmer, W.F., 1981. *Genética de las poblaciones humanas*. Ed. Omega, Barcelona.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. *Ecology*, 26: 297-302. <http://dx.doi.org/10.2307/1932409>

- Gower, J.C., 1971. A general Coefficient of Similarity and some of its properties. *Biometrics*, 27: 857-871. <http://dx.doi.org/10.2307/2528823>
- Hillis, D.M., 1984. Misuse and modification of Nei's genetic distance. *Systematic Zoology* 33:238-240.
- Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Société vaudoise des Sciences Naturelles*, 44: 223-270.
- Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27 (2): 209-220.
- Nei, M., 1972. Genetic Distance between Populations. *American Naturalist*, 106: 283-292. <http://dx.doi.org/10.1086/282771>
- Peakall, R., Smouse, P.E., 2006. GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, 6: 288-295.
- Peakall, R., Smouse, P.E., 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics*, 28: 2537-2539.
- Ringnér, M., 2008. What is principal component analysis? *Nature Biotechnology*, 26: 303-304. <https://doi.org/10.1038/nbt0308-303>
- Rogers, J.S., 1972. Measures of genetic similarity and genetic distance. *Studies in Genetics VII, University of Texas Publication*, 7213: 145-153
- Rogers, D., Tanimoto, T., 1960. A computer program for classifying Plants Science, 132, 1115-1118. <http://dx.doi.org/10.1126/science.132.3434.1115>
- Rohlf, F.J., 1998. NTSyS-p.c. Numerical taxonomy and multivariate analysis system (Version 2.0). Exeter Software Publishers Ltd., Setauket.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating relationships. University of Kansas Science Bulletin, 38: 1409-1448.

Sokal, RR, Rohlf, FJ., 1962. The comparison of dendrograms by objective methods. Taxon, 11: 33-40.

Soto, A.J., Ponzoni, I., Vázquez, G.E., 2006. Análisis numérico de diferentes criterios de similitud en algoritmos de clustering. Mecánica Computacional, Vol XXV, pp. 993-1011. Alberto Cardona, Norberto Nigro, Victorio Sonzogni, Mario Storti. (Eds.)

Sorensen, T., 1948. A Method of establishing groups of equal amplitude in Plant Sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons. Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab, 5: 1-34.

Ward, J.H. Jr., 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58: 236–244.

Wright, S., 1950. Genetical structure of populations. Nature, 166: 247-249.

Wright, S., 1978. Variability within and among Natural Populations. Evolution and the Genetics of Populations. Vol. 4. University of Chicago Press, Chicago.

## **Análisis de pruebas genéticas**

### **1. Pruebas genéticas**

Los árboles que se seleccionan para fines de mejora genética suelen seleccionarse por tener un buen fenotipo. Una vez seleccionados, su valor genético se prueba cruzándolos y su descendencia (progenie) se evalúa mediante pruebas genéticas (Zobel y Talbert, 1988).

Los objetivos de las pruebas genéticas son los siguientes:

- 1) Prueba de progenie. La mejor forma de evaluar el valor genético de los progenitores seleccionados es cultivando su progenie con el fin de poder estimar los valores de mejora de los parentales.
- 2) Estimación de los componentes de la varianza y de la heredabilidad. La elección de qué características mejorar depende bastante de la heredabilidad de estas.
- 3) Producción de una población base para las siguientes generaciones de selección y cruzamiento. Posiblemente una de las funciones más importantes de la prueba genética sea la de proporcionar una fuente de material a partir de la cual puedan hacerse las selecciones para la siguiente generación.
- 4) Estimación de la ganancia genética. Se trata de comparar el rendimiento relativo de las líneas mejoradas y no mejoradas en la misma prueba.

Ningún diseño de prueba genética puede cubrir todos los objetivos anteriores.

Una vez definidos los objetivos de un programa de pruebas genéticas, se debe elegir el diseño de apareamiento (policruzamientos, dialelos, cruzamientos jerárquicos, cruzamientos factoriales, etc.) que va a utilizarse para producir la población de la progenie (familias de medio hermanos, familias de hermanos completos) y el diseño experimental (bloques completos al azar, bloques incompletos, etc.) que se va a utilizar cuando la prueba se establezca en el campo.

Uno de los mayores problemas en la selección de genotipos con una alta productividad en diferentes ambientes (localidades, años) es la interacción genotipo x ambiente (GxE). Los nuevos genotipos además de tener una buena producción/rendimiento, deberán tener una buena estabilidad (capacidad de comportarse bien en una gran amplitud de condiciones ambientales) y adaptabilidad (capacidad de respuesta a la mejora del

ambiente, como por ejemplo al laboreo, abonado, riego, etc.). A mayor estabilidad y adaptabilidad genética de los genotipos, menor interacción GxE.

Varios métodos de análisis de adaptabilidad y estabilidad se han propuesto por diversos autores (Linn y Binns, 1988; Annicchiarico, 1992). Estos métodos asumen que los efectos genotípicos son fijos y según Resende (2007), esto es algo limitante en el análisis de experimentos no equilibrados (cuando los tratamientos son asignados a un número desigual de unidades experimentales), no ortogonales (cuando cada nivel de un factor no está presente en el experimento en combinación con cada nivel de otro factor y no se pueden investigar las interacciones) y con heterogeneidad de varianzas (cuando la varianza de los errores no es constante en todas las observaciones realizadas). Por otra parte, en los modelos mixtos, la predicción de los valores genotípicos de los genotipos por el método de la Mejor Predicción Linear Insesgada (Best Linear Unbiased Prediction, BLUP) y la estimación de los componentes de la varianza por el Método de Máxima Verosimilitud Restringida (Restricted Maximum Likelihood, REML), permiten la predicción de los efectos genéticos (asumidos aleatorios) libres de los efectos fijos del modelo (Henderson, 1975; Piepho et al., 2008). La evaluación de la significación de los efectos aleatorios en los modelos mixtos no se hace con la prueba F del análisis de varianza, sino con la razón de verosimilitud (LRT) que se obtiene en el análisis de desviación o devianza (cantidad de varianza explicada por el modelo) en un modelo lineal generalizado mixto o GLM (Resende, 2007). Los GLM son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes (McCullagh y Nelder, 1989). Para la selección simultánea de producción, estabilidad y adaptabilidad en genotipos, Resende (2007) propuso la media armónica del rendimiento relativo (%) de los valores genotípicos estimados (MHPRVG). Este método se ha probado útil con diferentes cultivos (Pinto Jr. et al., 2006; Verardi et al., 2009; Mendes et al., 2012; Oliveira y Bande, 2024).

## **2. Introducción al programa SELEGEN-REML/BLUP**

Procedimientos estadísticos avanzados, como los modelos mixtos (que combinan efectos aleatorios con efectos fijos) por medio de los métodos REML y BLUP están presentes en el software SELEGEN-REML/BLUP desarrollado por Resende en EMBRAPA, Brasil (Resende 2007; Resende, 2016). La ventaja de estos métodos es que

realizan un mejor manejo de los ensayos no equilibrados, para la obtención de parámetros estadísticos sin sesgo.

Este programa realiza la estimación de los componentes de varianza con el método REML y la predicción de los valores genéticos (valores de mejora) con el procedimiento BLUP. La estimación de los valores de mejora para cada genotipo a través del procedimiento BLUP incluye el valor genético del genotipo, libre de los efectos ambientales incluidos en el modelo.

En la página web: [https://www.ppestbio.ufv.br/?page\\_id=448](https://www.ppestbio.ufv.br/?page_id=448), se descarga el fichero, seleccionando al final de la página: Software Selegen RemlBlup, botón derecho del ratón y seleccione Guardar enlace como: Selegen\_20190424\_E.zip.

Una vez que se tenga el fichero .zip, se copia en el escritorio de nuestro ordenador y se crea una carpeta que se puede llamar Selegen, metiendo dicho fichero en su interior y con el botón derecho del ratón, se extraen todos los ficheros en esa carpeta.

Los datos tomados en campo se deben preparar en un archivo Excel. SELEGEN no acepta datos faltantes ni el uso de comas para los decimales.

Cuando no hay datos se puede colocar un “0”. SELEGEN tiene la opción de omitir los valores con “0”.

Una vez “limpios” los datos, se procede a ordenarlos como lo solicita SELEGEN, por ejemplo: Individuo Familia/Clon Bloque Parcela Árbol Variables

La Parcela es la interacción BloquexFamilia.

Matriz de datos organizados para la ejecución del programa SELEGEN (Resende et al., 2018):

	A	B	C	D	E	F	G	H	
1	Individuo	Familia	Bloque	Parcela	Arbol	Altura	Diametro	Volumen	
2	1	1	1	11	1	303	4	5	
3	2	1	1	11	2	440	8	28	
4	3	1	1	11	3	425	8	27	
5	4	1	1	11	4	352	6	13	
6	5	1	1	11	5	451	8	29	
7	6	1	2	21	1	435	8	28	
8	7	1	2	21	2	351	5	9	
9	8	1	2	21	3	438	7	21	
10	9	1	2	21	4	435	7	21	
11	10	1	2	21	5	436	7	21	

El archivo generado en Excel debe grabarse con la extensión “txt” (Texto MS-DOS)

Una vez creado el fichero de datos y guardado en formato “txt” se guarda en la carpeta de Selegen y se hace clic en el fichero: SELEGEN\_REMLBLUP\_W.exe.

Seleccionar “Si” para Permitir que la aplicación haga cambios en el equipo.

Inmediatamente se despliega una ventana:

The screenshot shows the SELEGEN software interface. At the top, the title bar reads "Procedimientos Matemáticos / Estadísticos / Genéticos". Below this, there are two columns of buttons. The left column contains buttons for: "Modelos Mistos: Delineamentos Experimentais / Materiais Genéticos / Ambientes" (highlighted with a dashed border), "Repetibilidade e Dados Longitudinais", "Dialélicos - Genitores não Aparentados", "Produtividade, Estabilidade e Adaptabilidade", "Análise Multivariada: Divergência Genética e Agrupamento", "Otimização da Seleção ( Endogamia e Ne )", "Melhoramento Animal", "Famílias F3 ou S1", "Clones Aparentados / Matriz A Completa", "Interação com Locais e Anos", "Análise de Variância / REML", "Autocorrelação Espacial", and "Dados Categóricos". The right column contains buttons for: "Modelos Mistos com Covariável", "Fatoriais Interpopulacionais", "Dialélicos - Genitores Aparentados", "Índice de Seleção", "Modelos Mistos - Tratamentos de Efeitos Fixos", "Correlações Genéticas", "Genética de Populações", "Sistema Reprodutivo Misto", "Várias Populações ou Procedências", "Modelos Genéticos", "Espécies Perenes Autógamas: Café Árábica, Pêssego, Anuais", "Interação Planta x Patógeno", and "Estatística Geral". At the bottom left, there are radio buttons for "Convergência" with options "Deviance" and "Herdabilidades" (selected). At the bottom right, there is a text input field for "Fator de aceleração da convergência" with the value "0.0".

Se selecciona dentro de Procedimientos Matemáticos/Estadísticos/Genéticos de los disponibles en la página, el que nos interese.

Por ejemplo:

Selección de Modelos mistos: Delineamentos experimentais/Materiais genéticos/Ambientes

Arquivo

Selecciono el archivo en formato “.txt”

Abrir

Delineamento Experimental: Blocos ao Acaso

Modelo

Por ejemplo, selecciono 001. Blocos ao acaso, progênies de meios irmãos, várias plantas por parcela (Bloques completos, progenies de medios hermanos, varias plantas por parcela)

Se escribe el número de variables que tiene el fichero “.txt”.

Se indica la variable que se desea analizar.

El programa analiza una variable de cada vez

Finalmente se ejecuta el programa, presionando clic en Ejecutar.

Obtengo los resultados de la primera variable en la pantalla, creando el programa siete ficheros en el directorio Selegen:

Fichero “.Res”:  
presenta para cada variable los componentes de la varianza, los valores genéticos predichos (estimados), el progreso genético con selección y el tamaño efectivo poblacional.

Fichero “.Fam”:  
presenta para cada variable los valores genéticos predichos y sus intervalos de confianza.

Fichero “.Dev”:  
presenta para cada variable el valor máximo de la función de verosimilitud restringida, la desvianza (devianza), los residuos y el cuadrado medio referente a los factores de efectos fijos, el cual se puede utilizar para hacer el test F para el factor referido de efectos fijos.

Fichero “.Efe”:  
presenta para cada variable todos los efectos ajustados por el modelo utilizado (Bloque, Tratamiento, BloquexTratamiento o Parcela).

Fichero “.Het”:  
presenta para cada variable la varianza residual dentro de cada tratamiento y las heredabilidad individual válida para cada tratamiento

Los ficheros “.Err” y “.Dad” no se utilizan.

**3. Análisis de ensayos de progenies de medios hermanos o de polinización abierta (especies alógamas). Diseño en bloques completos al azar con una planta por parcela. Evaluación en una sola localidad.**

**Fichero: DatapinasterHS1loca-1planta.xls (Zas et al., 2004)**

El fichero con datos de *Pinus pinaster* tiene la siguiente estructura: Diseño en bloques completos al azar, 10 familias de medios hermanos de una sola procedencia, 2 repeticiones y 1 observación/parcela.

VARIABLES medidas a los 8 años: variable 1 = altura (cm), variable 2 = diámetro a 1,3 m (cm) y variable 3 = volumen (diámetro<sup>2</sup> x altura/1000) en dm<sup>3</sup>.

Las columnas se configuran en formato número.

El orden de las columnas debe de ser el siguiente:

Individuo Familia Repetición Observación/Parcela variable1 variable2 variable3

**Objetivos: Determinar los componentes de la varianza y los valores genotípicos de la variable volumen.**

Una vez creado el fichero de datos en Excel y guardado en formato texto (MS-DOS) se guarda en la carpeta de Selegen y se hace clic en el fichero:

SELEGEN\_REMLBLUP\_W.exe

Clic en Si

Selección de Análise de Variãncia/REML.

Archivo

Selecciono el archivo: **DatapinasterHS1loca-1planta.txt**

Abrir

Modelo

Selecciono 095. ANOVA e REML: Blocos ao acaso, progênies de meios irmãos, uma planta por parcela (Bloques completos, progenies de medios hermanos, una planta por parcela)

$$y = Xr + Za + e$$

y = vector de datos fenotípicos

r = vector efecto repetición (fijo)

a = vector efectos genéticos aditivos individuales (aleatorio)

e = vector de error o residual (aleatorios)

X y Z son las matrices de incidencia para los efectos mencionados

Número de Variáveis: 3

Ordem da Variável a ser Analisada: 3

Valores Iniciais dos Parâmetros h2 0.1

REML/BLUP

Considerar Zeros

Não

Taxa de Erro 0.00001

SELEGEN-REML/BLUP: Sistema Estatístico e Seleção Genética Computadorizada

Executar Voltar Sair

Arquivo C:\Users\Oliveira\Desktop\Selegen\Selegen\_20190424\_E\DatapinasterHS1loca-1planta.txt

Modelo 095. ANOVA e REML: Blocos ao acaso, progênies de meios irmãos, uma planta por parcela

Número de Variáveis: 3

Ordem da Variável a ser Analisada: 3

Valores Iniciais dos Parâmetros h2 0.1

REML/BLUP  
 BLUP

Considerar Zeros  
 Sim  
 Não

Taxa de Erro 0.00001

Clic en Executar

## Los resultados del análisis de la variable volumen son:

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Viçosa      Viçosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 2016

Arquivo      : C:\Users\Usuario\Desktop\Selegen\Selegen_20190424_E\DataspinasterHS1loca-iplanta.txt
Modelo       : 95
Numero de Variaveis : 3
Variavel Analisada : 3
N. Dados Lidos   : 20
Zeros significativos : Nao
Deviance      : 101.26
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63          2.5% = 3.84
5% = 3.84          5% = 2.71
10% = 2.71
Data : 04/12/2023
Hora : 12:04:00
```

## Componentes de la varianza

### 1. Componentes de Variância ( REML Individual )

```
Va      =      0.633529
Ve      =      78.443404
Vf      =      79.076933
h2a     =      0.008012 +- 0.1132
h2mp    =      0.003998
Acprog  =      0.063228
h2ad    =      0.006021
CVgi%   =      3.940324
CVgp%   =      1.970162
CVe%    =      43.978274
CVr     =      0.044799
Média geral = 20.200000
```

	FV	GL	SQ	QM	F
Ef. Fixo	1	96.8000	96.8000	1.2266	
Progenie	9	713.1178	79.2353	1.0040	
Residuo	9	710.2670	78.9186	-	

### Medias dos Efeitos Fixos

Ef. Fixo	Valor
1	18.0000
2	22.4000

## Notas:

Va = varianza genética aditiva.

Ve = varianza residual (ambiental).

Vf = varianza fenotípica total.

h2a = h2 = heredabilidad individual.

h2mp = heredabilidad promedio de la familia

Acprog = Precisión de la selección de familias.

h2ad = heredabilidad aditiva dentro de familia.

CVgi% = coeficiente de variación genética individual.

CVgp% = coeficiente de variación genotípica entre familias.

CVe% = coeficiente de variación residual.

CVr = CVgp/CVe = coeficiente de variación relativa.

Média geral = Media general del experimento.

## Valores genotípicos

2. Componentes de Média ( BLUP Individual )

Seleção de Indivíduos

Ordem	Bloco	Familia	f	a	u+a	Ganho	Nova Média	Ne	d	g
1	1	10	33.0000	0.1194	20.3194	0.1194	20.3194	1.0000	0.0601	0.1794
2	1	3	29.0000	0.0993	20.2993	0.1093	20.3093	2.0000	0.0440	0.1433
3	2	9	32.0000	0.0729	20.2729	0.0972	20.2972	3.0000	0.0385	0.1114
4	2	3	28.0000	0.0667	20.2667	0.0896	20.2896	3.4909	0.0223	0.0891
5	2	7	33.0000	0.0610	20.2610	0.0839	20.2839	4.4944	0.0426	0.1036
6	1	2	27.0000	0.0534	20.2534	0.0788	20.2788	5.4962	0.0361	0.0895
7	2	10	22.0000	0.0266	20.2266	0.0713	20.2713	6.0681	-0.0017	0.0249
8	1	5	21.0000	0.0212	20.2212	0.0651	20.2651	7.0588	0.0120	0.0333
9	2	1	28.0000	0.0190	20.2190	0.0600	20.2600	8.0521	0.0225	0.0415
10	1	9	16.0000	0.0031	20.2031	0.0543	20.2543	8.6420	-0.0081	-0.0050
11	1	6	21.0000	-0.0006	20.1994	0.0493	20.2493	9.6284	0.0121	0.0115
12	2	5	21.0000	-0.0052	20.1948	0.0447	20.2447	10.2335	-0.0056	-0.0109
13	2	4	21.0000	-0.0291	20.1709	0.0390	20.2390	11.2147	-0.0055	-0.0347
14	1	8	13.0000	-0.0528	20.1472	0.0325	20.2325	12.1992	-0.0200	-0.0728
15	2	2	13.0000	-0.0574	20.1426	0.0265	20.2265	12.8063	-0.0377	-0.0951
16	2	8	16.0000	-0.0612	20.1388	0.0210	20.2210	13.4266	-0.0256	-0.0868
17	1	4	9.0000	-0.0749	20.1251	0.0154	20.2154	14.0576	-0.0360	-0.1109
18	1	7	6.0000	-0.0750	20.1250	0.0104	20.2104	14.6976	-0.0482	-0.1232
19	1	1	5.0000	-0.0930	20.1070	0.0049	20.2049	15.3454	-0.0521	-0.1451
20	2	6	10.0000	-0.0934	20.1066	0.0000	20.2000	16.0000	-0.0497	-0.1431

### Notas:

f = Valores fenotípicos observados

a = Efectos aditivos

u = Media general

u + a = Valores genéticos aditivos

Ganho = Ganancia genética esperada sobre la Media general

Nova Média = Promedio mejorado (Media general + Ganancia genética)

Ne = Tamaño efectivo poblacional

d = Efectos de dominancia

g = Efectos genotípicos totales (a+d)

El fichero “.Res” permite ver el ranking de los individuos y de las familias con base en su valor genotípico (Media general + a + d).

En este caso al tratarse de familias de medios hermanos el ranking se basa en el valor genético aditivo estimado = “a” + “u”.

La columna “Ordem” organiza el ranking de mayor a menor.

La columna “Ganho” es la ganancia genética esperada sobre la Media general si se selecciona ese individuo como progenitor, junto con todos los que le preceden en el ranking.

La columna “Nova Média” indica cual es el nuevo valor promedio estimado de la plantación si se utiliza ese individuo como progenitor y todos los que le preceden en el ranking. “Ne” es el tamaño efectivo de la población si se selecciona a ese individuo junto con todos los que le preceden en el ranking. El tamaño efectivo poblacional es una función del número de individuos y su grado de parentesco, por lo que a mayor relación genética (parentesco), se requerirá una mayor cantidad de individuos para alcanzar un “Ne” superior.

La columna “d” es la estimación del efecto genético no aditivo o de dominancia. Finalmente, la columna “g” es el efecto genotípico total estimado (“a” + “d”) de cada individuo.

**4. Análisis de ensayos de progenies de medios hermanos o de polinización abierta. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en una sola localidad.**

**Fichero: DatapinasterHS1loca-plantas.xls (Zas et al., 2004)**

El fichero con datos de *Pinus pinaster* tiene la siguiente estructura: Diseño en bloques completos al azar, 10 familias de medios hermanos de una sola procedencia, 10 bloques, 5 plantas/bloque.

VARIABLES medidas a los 8 años: variable 1 = altura (cm), variable 2 = diámetro a 1,3 m (cm) y variable 3 = volumen (diámetro<sup>2</sup> x altura/1000) en dm<sup>3</sup>.

Las columnas se configuran en formato número.

El orden de las columnas debe de ser el siguiente:

Individuo Familia Bloque Parcela Árbol variable1 variable2 variable3

Parcela es la interacción BloquexFamilia

**Objetivos: Determinar los componentes de la varianza, los valores genotípicos y los intervalos de confianza, las correlaciones genéticas y fenotípicas entre las variables estudiadas.**

Una vez creado el fichero de datos en Excel y guardado en formato texto (MS-DOS) se guarda en la carpeta de Selegen y se hace clic en el fichero:

SELEGEN\_REMLBLUP\_W.exe

Clic en Si

Se abre una ventana del programa.

Selección de Modelos mixtos: Delineamientos experimentais/Materiais genéticos/Ambientes

Arquivo

Selecciono el archivo: **DatapinasterHS1loca-plantas.txt**

Abrir

Delineamento Experimental: Blocos ao Acaso

Modelo

Selecciono 001. Blocos ao acaso, progênies de meios irmãos, várias plantas por parcela (Bloques completos, progenies de medios hermanos, varias plantas por parcela)

$$y = Xr + Za + Wp + e$$

y = vector de datos fenotípicos

r = vector efecto repetición (efecto fijo)

a = vector efectos genéticos aditivos individuales (efecto aleatorio)

p = vector efecto parcela (efecto aleatorio)

e = vector de error o residual (efecto aleatorio)

X, Z y W son las matrices de incidencia para los efectos mencionados

Número de Variáveis: 3

Ordem da Variável a ser Analisada: 1

Valores Iniciais dos Parâmetros h2 0.1

REML/BLUP

Considerar Zeros:

Não

Taxa de Erro 0.00001

Clic en Executar

Resultados de la primera variable:

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genetica Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Viçosa   Viçosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 201

Arquivo      : C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_E\DatapinasterHSiloca-plantas.txt
Modelo       : 1
Numero de Variaveis : 3
Variavel Analisada : 1
N. Dados Lidos : 500
Zeros significativos : Nao
Deviance     : 4453.54
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63          2.5% = 3.84
5% = 3.84          5% = 2.71
10% = 2.71
Data : 06/05/2024
Hora : 11:00:22
```

## Componentes de la varianza

### 1. Componentes de Variância ( REML Individual )

Va	=	1852.404147	
Vparc	=	1266.803074	
Ve	=	914.551548	
Vf	=	4033.758769	
h2a	=	0.459225	+ - 0.1714
h2aj	=	0.669474	
c2parc	=	0.314050	
h2mp	=	0.728308	
Acprog	=	0.853410	
h2ad	=	0.603034	
CVgi%	=	9.831547	
CVgp%	=	4.915774	
CVe%	=	9.494505	
CVr	=	0.517749	
PEV	=	125.820665	
SEP	=	11.216981	
Média geral	=	437.770000	

## Notas:

Va = variância genética aditiva.

V<sub>parc</sub> = variância ambiental entre parcelas.

Ve = variância residual (ambiental).

Vf = variância fenotípica total.

h<sub>2a</sub> = h<sub>2</sub> = heredabilidade individual.

h<sub>2aj</sub> = heredabilidade individual de parcela.

c<sub>2parc</sub> = c<sub>2</sub> = coeficiente de determinação de los efectos de parcela.

h<sub>2mp</sub> = heredabilidade promedio de la familia

Acprog = Precisión de la selección de familias.

h<sub>2ad</sub> = heredabilidade aditiva dentro de familia.

CV<sub>gi</sub>% = coeficiente de variación genética individual.

CV<sub>gp</sub>% = coeficiente de variación genotípica entre familias.

CV<sub>e</sub>% = coeficiente de variación residual.

CV<sub>r</sub> = CV<sub>gp</sub>/CV<sub>e</sub> = coeficiente de variación relativa.

PEV = variância del error de predicción de los valores genotípicos de familia.

SEP = desviación estándar del valor genotípico predicho de familia.

Média geral = Media general del experimento.

## Valores genotípicos

2. Componentes de Média ( BLUP Individual )

Seleção de Individuos

Ordem	Bloco	Familia	Árvore	f	a	u+a	Ganho	Nova Média	Ne	d	g
1	9	8	3	633.0000	107.0950	544.8650	107.0950	544.8650	1.0000	46.0270	153.1220
2	7	7	3	615.0000	87.7733	525.5433	97.4341	535.2041	2.0000	47.3043	135.0776
3	5	8	3	592.0000	81.5337	519.3037	92.1340	529.9040	2.4828	28.9861	110.5198
4	1	7	4	609.0000	80.3234	518.0934	89.1813	526.9513	3.2000	42.3377	122.6611
5	7	2	5	610.0000	76.5561	514.3261	86.6563	524.4263	4.1096	51.1199	127.6760
6	10	8	1	610.0000	75.3222	513.0922	84.7673	522.5373	4.3636	24.8452	100.1674
7	1	7	5	599.0000	74.2930	512.0630	83.2709	521.0409	4.7419	38.3175	112.6105
8	4	8	4	530.0000	72.2942	510.0642	81.8988	519.6688	4.8917	22.8265	95.1206
9	3	8	4	513.0000	71.5432	509.3132	80.7482	518.5182	4.9091	22.3258	93.8690
10	3	1	4	517.0000	70.2555	508.0255	79.6989	517.4689	5.7416	44.2199	114.4754
11	5	8	5	573.0000	70.0760	507.8460	78.8241	516.5941	5.6553	21.3477	91.4237
12	10	7	5	620.0000	68.1296	505.8996	77.9329	515.7029	6.0000	34.2085	102.3381
13	4	8	1	523.0000	68.0729	505.8429	77.1745	514.9445	5.9168	20.0123	88.0852
14	1	8	3	551.0000	66.8836	504.6536	76.4394	514.2094	5.8074	19.2194	86.1030
15	7	8	4	512.0000	63.7349	501.5049	75.5924	513.3624	5.6872	17.1203	80.8552
16	3	3	4	579.0000	63.5386	501.3086	74.8391	512.6091	6.3920	46.3259	109.8645
17	10	8	3	590.0000	63.2615	501.0315	74.1580	511.9280	6.2385	16.8047	80.0662
18	4	8	2	515.0000	63.2486	501.0186	73.5520	511.3220	6.0902	16.7961	80.0448
19	6	2	1	508.0000	60.2543	498.0243	72.8521	510.6221	6.6330	40.2521	100.5064
20	3	7	5	441.0000	59.6517	497.4217	72.1921	509.9621	6.9565	28.5566	88.2082
21	6	8	4	477.0000	59.3592	497.1292	71.5810	509.3510	6.7925	14.2032	73.5624
22	1	8	5	538.0000	59.0441	496.8141	71.0111	508.7811	6.6347	13.9931	73.0372
23	10	8	2	580.0000	57.2312	495.0012	70.4120	508.1820	6.4848	12.7845	70.0157
24	2	8	5	491.0000	56.5643	494.3343	69.8350	507.6050	6.3436	12.3399	68.9042
25	9	8	5	549.0000	56.4401	494.2101	69.2992	507.0692	6.2112	12.2571	68.6972
26	1	8	2	533.0000	56.0290	493.7990	68.7888	506.5588	6.0873	11.9830	68.0120
27	3	10	3	505.0000	54.4441	492.2141	68.2575	506.0275	6.5988	34.6598	89.1039
28	9	1	5	559.0000	53.5528	491.3228	67.7324	505.5024	7.0209	33.0848	86.6375
29	3	10	2	502.0000	52.6350	490.4050	67.2118	504.9818	7.4529	33.4537	86.0888
30	5	8	2	543.0000	51.9850	489.7550	66.7042	504.4742	7.2816	9.2870	61.2720

Si se desea seleccionar la familia (la madre), aparece un ranking similar con todas las familias participantes bajo la columna “Genitor”.

Aparece el efecto genético aditivo estimado “a”, la ganancia genética esperada (Ganho) y el nuevo valor promedio esperado (Nova Média) si se utiliza esa familia y todas las que la precedan en el ranking. Por tratarse de familias de medios hermanos, el ranking genético está organizado de mayor a menor, por el valor genético aditivo estimado “a”.

### Selección de progenitores (familia)

#### Seleção de Genitores

Ordem	Genitor	a	Ganho	Nova Média
1	8	76.1089	76.1089	513.8789
2	7	33.6336	54.8712	492.6412
3	1	7.8512	39.1979	476.9679
4	10	4.9088	30.6256	468.3956
5	2	-0.2476	24.4510	462.2210
6	6	-1.2381	20.1695	457.9395
7	4	-10.2984	15.8169	453.5869
8	3	-11.9007	12.3522	450.1222
9	9	-48.7534	5.5627	443.3327
10	5	-50.0643	0.0000	437.7700

Si se desea comparar la posición en el ranking que alcanzarían las madres y sus progenies, aparece una lista completa con todos los individuos del ensayo y sus progenitoras, que se distinguen porque SELEGEN los ubicará en el “Bloco 0”, por no estar presente físicamente dentro del ensayo de progenie. Esta lista conjunta puede ser de utilidad en aquellos casos donde las madres aún existen en el programa y se desee analizar si aportaría mayor ganancia genética seleccionar a la madre que a su propia descendencia.

## Selección con sobreposición de generaciones (ranking unificado de las madres y su progenie)

Seleção com Sobreposição de Gerações

Ordem	Bloco	Familia	Árvore	a	Ganho	Nova Média
1	9	8	3	107.0950	107.0950	544.8650
2	7	7	3	87.7733	97.4341	535.2041
3	5	8	3	81.5337	92.1340	529.9040
4	1	7	4	80.3234	89.1813	526.9513
5	7	2	5	76.5561	86.6563	524.4263
6		0 8	0	76.1089	84.8984	522.6684
7	10	8	1	75.3222	83.5303	521.3003
8	1	7	5	74.2930	82.3757	520.1457
9	4	8	4	72.2942	81.2555	519.0255
10	3	8	4	71.5432	80.2843	518.0543
11	3	1	4	70.2555	79.3726	517.1426
12	5	8	5	70.0760	78.5979	516.3679
13	10	7	5	68.1296	77.7926	515.5626
14	4	8	1	68.0729	77.0983	514.8683
15	1	8	3	66.8836	76.4174	514.1874
16	7	8	4	63.7349	75.6247	513.3947
17	3	3	4	63.5386	74.9138	512.6838
18	10	8	3	63.2615	74.2664	512.0364
19	4	8	2	63.2486	73.6865	511.4565
20	6	2	1	60.2543	73.0149	510.7849
21	3	7	5	59.6517	72.3786	510.1486
22	6	8	4	59.3592	71.7868	509.5568
23	1	8	5	59.0441	71.2328	509.0028
24	10	8	2	57.2312	70.6494	508.4194
25	2	8	5	56.5643	70.0860	507.8560
26	9	8	5	56.4401	69.5611	507.3311
27	1	8	2	56.0290	69.0599	506.8299
28	3	10	3	54.4441	68.5379	506.3079
29	9	1	5	53.5528	68.0212	505.7912
30	3	10	2	52.6350	67.5083	505.2783
31	5	8	2	51.9850	67.0076	504.7776
32	4	7	2	51.5384	66.5242	504.2942

## Valores genéticos (VG) e Intervalos de confianza (LIIC y LSIC)

En el archivo “.Fam” situado en el directorio de Selegen, aparecen los límites de confianza para cada familia estimados mediante la expresión  $(u+g) \pm t \text{ SEP}$ , siendo  $t$  la  $t$  de Student (para muestras  $< 30$ ) y SEP la desviación estándar del valor genotípico estimado de la familia ( $u + g = u + a + d$ ).

Se muestra la salida del archivo “.Fam”, donde LIIC y LSIC son el Límite Inferior del Intervalo de confianza y el Límite Superior del Intervalo de confianza respectivamente para la variable 1 (altura).

1	445.6212
2	437.5224
3	425.8693
4	427.4716
5	387.7057
6	436.5319
7	471.4036
8	513.8789
9	389.0166
10	442.6788

Ordem	Genotipo	VG	Acuacia	LIIC	LSIC	Med	Fenot
1	8	475.8244	0.8096	451.0672	500.5816	490.0200	
2	7	454.5868	0.8096	429.8296	479.3440	460.8600	
3	1	441.6956	0.8096	416.9384	466.4528	443.1600	
4	10	440.2244	0.8096	415.4672	464.9816	441.1400	
5	2	437.6462	0.8096	412.8890	462.4034	437.6000	
6	6	437.1509	0.8096	412.3937	461.9081	436.9200	
7	4	432.6208	0.8096	407.8636	457.3780	430.7000	
8	3	431.8197	0.8096	407.0625	456.5769	429.6000	
9	9	413.3933	0.8096	388.6361	438.1505	404.3000	
10	5	412.7378	0.8096	387.9806	437.4950	403.4000	

## Notas:

VG = Valores genéticos estimados (“Valores BLUP”)

Accuracia = Precisión de los valores genéticos estimados

Med Fenot = Media fenotípica de cada genotipo

Lo mismo se haría con las variables 2 y 3, acumulándose los ficheros de resultados en la carpeta Selegen, con las denominaciones \_V01, \_V02 y \_V03 para las variables 1, 2 y 3 respectivamente.

## Correlaciones genéticas

Una vez corridas las tres variables, minimizar la ventana del modelo anterior y clicar en “Correlacoes genéticas” (modelo 102) en la segunda columna de la pantalla principal de SELEGEN.

## Clic en Executar.

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Viçosa      Viçosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 2016.

Arquivo      : C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_E\DatapinasterLalinoriginal2 Jose Alberto Oliveira, Oviedo, Enero23.txt
Modelo       : 102
Numero de Variaveis : 1
Variavel Analisada : 1
N. Dados Lidos : 0
Zeros significativos : Nao
Deviance     : 0.00
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63          2.5% = 3.84
5% = 3.84          5% = 2.71
10% = 2.71
Data : 01/12/2023
Hora : 13:00:47
```

### Matriz de Covariância Genética

Variavel	Var.	1 Var.	2 Var.	3
1	1349.1253	22.7589	222.3684	
2	22.7589	0.5672	4.9339	
3	222.3684	4.9339	45.4253	

### Matriz de Correlação Genética

Variavel	Var.	1 Var.	2 Var.	3
1	1.0000	0.8227	0.8983	
2	0.8227	1.0000	0.9720	
3	0.8983	0.9720	1.0000	

## Correlaciones fenotípicas

Minimizar la ventana anterior

Clic en “Estatística Geral” (modelo 105) en la segunda columna de la pantalla principal de SELEGEN.

Clic en Executar

El programa pregunta lo siguiente:

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS ò 200
Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Vicosa   Vicosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citacao: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechn
ology 16: 330-339, 2016.
Numero de variaveis :
```

Número de variables: 3 Enter

```
C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_E\Selegen_W.Exe
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS ò 200
Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Vicosa   Vicosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citacao: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechn
ology 16: 330-339, 2016.
Numero de variaveis :
3
Iniciar a partir de qual coluna
```

Iniciar a partir de qual coluna: 6 Enter

Si el programa vuelve a la pantalla principal, clicar en Estatística Geral otra vez

## Aparece lo siguiente:

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
Universidade Federal de Vigosa Vigosa - MG
marcos.deon@gmail.com
marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 2016.

Arquivo : C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_F\DatapinasterLalinoriginal2 Jose Alberto Oliveira, Oviedo, Enero23.txt
Modelo : 105
Numero de Variaveis : 1
Variavel Analisada : 1
N. Dados Lidos : 0
Zeros significativos : Nao
Deviance : 0.00
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63 2.5% = 3.84
5% = 3.84 5% = 2.71
10% = 2.71
Data : 01/12/2023
Hora : 13:05:20

1. Valores Observados

Genotipo Var01 Var02 Var03
1 303.0000 4.0000 5.0000
2 440.0000 6.0000 28.0000
3 425.0000 8.0000 27.0000
4 352.0000 6.0000 13.0000
5 451.0000 9.0000 29.0000
6 437.0000 7.0000 21.0000
7 420.0000 7.0000 21.0000
8 490.0000 8.0000 31.0000
9 353.0000 5.0000 9.0000
10 367.0000 5.0000 9.0000
11 369.0000 7.0000 18.0000

2. Estatisticas Basicas

Media 437.7700 7.2420 26.1660
Variancia 5080.6584 3.6267 288.4353
Desvio 71.2787 1.9044 16.9834
CV % 16.2822 26.2964 64.9063
Maximo 641.0000 14.0000 124.0000
Minimo 198.0000 2.0000 1.0000

Assim(a3) -0.0144 0.0802 1.5573
Desvio(a3) 0.1095 0.1095 0.1095
t (a3) -0.1313 0.7319 14.2166
Infer (a3) Simetrica Simetrica Assim.Pos

Curtos(a4) 0.7005 0.3193 3.8850
Desvio(a4) 0.2191 0.2191 0.2191
t (a4) 3.1975 1.4575 17.7327
Infer (a4) Leptocurt Mesocurt Leptocurt

3. Matriz de Covariancia

Variavel Var01 Var02 Var03
1 5080.6584 111.7151 1019.0262
2 111.7151 3.6267 30.5129
3 1019.0262 30.5129 288.4353

4. Matriz de Correlacao

Variavel Var01 Var02 Var03
1 1.0000 0.8230 0.8418
2 0.8230 1.0000 0.9434
3 0.8418 0.9434 1.0000
```

Las correlaciones fenotípicas entre las tres variables aparecen en el punto 4. Matriz de Correlacao.

**5. Análisis de ensayos de progenies de medios hermanos o de polinización abierta. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en varias localidades.**

**Fichero: DatapinasterHS4loca-plantas.xls (Zas et al., 2004)**

El fichero con datos de *Pinus pinaster* tiene la siguiente estructura: Diseño en bloques completos al azar, 10 familias de medios hermanos de una sola procedencia, 4 localidades, 2 repeticiones/localidad (repeticiones 1 y 2 en la localidad 1, repeticiones 3 y 4 en la localidad 2, repeticiones 5 y 6 en la localidad 3 y repeticiones 7 y 8 en la localidad 4) y 2 árboles/parcela.

VARIABLES medidas a los 8 años: variable 1 = altura (cm), variable 2 = diámetro a 1,3 m (cm) y variable 3 = volumen (diámetro<sup>2</sup> x altura/1000) en dm<sup>3</sup>.

Las columnas se configuran en formato número.

El orden de las columnas debe de ser el siguiente:

Individuo Familia Repetición Parcela Interacción Árbol variable1 variable2 variable3

La Interacción es LocalidadxFamilia.

**Objetivos: Determinar los componentes de la varianza, los valores genotípicos y el análisis de la desvianza considerando el factor familia y la interacción Localidad x Familias aleatorios para la variable altura.**

Una vez creado el fichero de datos en Excel y guardado en formato texto (MS-DOS) se guarda en la carpeta de Selegen y se hace clic en el fichero:

SELEGEN\_REMLBLUP\_W.exe

Clic en Si

Se abre una ventana del programa.

Selección de Modelos mixtos: Delineamientos experimentais/Materiais genéticos/Ambientes

Arquivo

Selecciono el archivo: **DatapinasterHS4loca-plantas.txt**

Abrir

Delineamento Experimental: Blocos ao Acaso

Modelo

Selecciono 004. Blocos ao acaso, progênies de meios irmãos, várias plantas por parcela, vários locais (Bloques completos, progenies de medios hermanos, varias plantas por parcela, varias localidades)

$$y = Xr + Za + Wp + Ti + e$$

y = vector de datos fenotípicos

r = vector efecto repetición (fijo)

a = vector efectos genéticos aditivos individuales del efecto familia (aleatorio)

p = vector efecto parcela (aleatorio)

i = vector de los efectos de interacción familia x localidad (aleatorios)

e = vector de error o residual (aleatorios)

X, Z, W y T son las matrices de incidencia para los efectos mencionados

Nº de variables: 3

Orden de variables a ser analizadas: 1.

Las demás opciones se dejan según aparecen en el modelo.

Clic en Executar

Se obtienen los resultados de la variable 1 (altura) en la pantalla.

Si al hacer los cálculos desaparecen los resultados de la pantalla es debido a que se minimiza. Para volver a ver los resultados clic en el icono de Selegen minimizado y se selecciona Modelos mistos: Delineamentos experimentais/Materiais genéticos/Ambientes

Aparecen los siguientes resultados:

```

SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Viçosa   Viçosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 201

Arquivo       : C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_E\DatapinasterHS4loca-plantas.txt
Modelo        : 4
Numero de Variaveis : 3
Variavel Analisada : 1
N. Dados Lidos   : 160
Zeros significativos : Nao
Deviance       : 1428.81
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63           2.5% = 3.84
5% = 3.84           5% = 2.71
10% = 2.71
Data : 06/05/2024
Hora : 12:08:04

```

1. Componentes de Variância ( REML Individual )

```

Va      = 19.656652
Vparc   = 2043.496953
Vint    = 14.068265
Ve      = 2338.263890
Vf      = 4415.485760
h2a     = 0.004452 +- 0.0298
c2parc  = 0.462802
c2int   = 0.003186
h2mp    = 0.011959
Acprog  = 0.109355
h2ad    = 0.006265
rgloc   = 0.258880
PEV     = 4.855396
SEP     = 2.203496
Média geral = 436.656250

```

2. Componentes de Média ( BLUP Individual )

Seleção de Individuos

Ordem	Bloco	Familia	Árvore	f	a	u+a	Ganho	Nova Média	Ne	d	g
1	1	7	2	517.0000	0.7589	437.4152	0.7589	437.4152	1.0000	0.3704	1.1293
2	5	3	1	610.0000	0.7321	437.3884	0.7455	437.4018	2.0000	0.3680	1.1001
3	6	7	2	579.0000	0.6989	437.3551	0.7300	437.3862	2.4828	0.3303	1.0292
4	4	3	2	434.0000	0.6831	437.3393	0.7182	437.3745	3.2000	0.3353	1.0184
5	3	3	2	620.0000	0.6497	437.3060	0.7045	437.3608	3.5088	0.3130	0.9627
6	3	7	1	490.0000	0.5267	437.1830	0.6749	437.3312	4.0000	0.2156	0.7423
7	4	7	2	487.0000	0.5238	437.1800	0.6533	437.3096	4.2151	0.2136	0.7374
8	3	4	1	605.0000	0.5029	437.1591	0.6345	437.2908	4.8917	0.2311	0.7340
9	5	7	1	511.0000	0.4839	437.1402	0.6178	437.2740	4.9091	0.1870	0.6709
10	1	8	2	455.0000	0.4574	437.1136	0.6017	437.2580	5.7416	0.1594	0.6168
11	2	7	2	475.0000	0.4479	437.1041	0.5877	437.2440	5.6553	0.1630	0.6108
12	8	8	2	510.0000	0.4459	437.1022	0.5759	437.2322	6.3529	0.1517	0.5976
13	3	5	2	590.0000	0.4386	437.0949	0.5654	437.2216	7.1686	0.2457	0.6844
14	5	8	2	530.0000	0.4217	437.0779	0.5551	437.2114	7.6712	0.1355	0.5572
15	4	8	1	508.0000	0.4214	437.0777	0.5462	437.2024	8.0000	0.1354	0.5568
16	4	8	2	508.0000	0.4214	437.0777	0.5384	437.1946	8.1789	0.1354	0.5568

#### Seleção de Genitores

Ordem Genitor	a	Ganho Nova Média	
1 8	0.4367	437.0929	
2 7	0.4067	437.0780	
3 3	0.3603	437.0575	
4 4	0.3124	437.0353	
5 5	0.1401	436.9875	
6 2	0.0906	436.9474	
7 6	-0.1056	436.8907	
8 1	-0.3558	436.8169	
9 9	-0.4097	436.7535	
10 10	-0.8756	0.0000	436.6562

#### Seleção com Sobreposição de Gerações

Ordem	Bloco	Familia	Árvore	a	Ganho Nova Média
1	1	7	2	0.7589	437.4152
2	5	3	1	0.7321	437.4018
3	6	7	2	0.6989	437.3862
4	4	3	2	0.6831	437.3745
5	3	3	2	0.6497	437.3608
6	3	7	1	0.5267	437.3312
7	4	7	2	0.5238	437.3096

### Análisis de la desvianza

Se vuelve a la pantalla principal del programa SELEGEN\_REML/BLUP y se corre el programa ajustando el modelo.

#### Pasos:

1. Ajustar el modelo sin el efecto Familia  $\zeta(h^2=0, c^2=0.1 \text{ y } c^{21}=0.1)$ ?

Clic en Executar

Anotar la desvianza: 1428.81.

2. Ajustar el modelo sin el efecto Familia x Localidad  $\zeta(h^2=0.1, c^2=0 \text{ y } c^{21}=0.1)$ ?

Clic en Executar

Anotar la desvianza: 1447.40.

3. Ajustar el modelo sin el efecto Parcela  $\zeta(h^2=0.1, c^2=0,1 \text{ y } c^{21}=0.0)$ ?

Clic en Executar

Anotar la desvianza: 1428.81.

4. Ajustar el modelo completo ( $h^2=0.1, c^2=0.1 \text{ y } c^{21}=0.1$ ).

Anotar la desvianza: 1428.81.

5. LRT para el efecto Familia =  $1428.81-1428.81 = 0.00$

6. LRT para el efecto Parcela =  $14128.81-1428.81 = 0.00$

7. LRT para el efecto Localidad x Familia =  $1447.4-1428.81 = 18.59$

8. Significación de los efectos aleatorios del modelo

En el archivo “.Res” se obtienen los valores del Qui-cuadrado (LRT Chi-cuadrado) tabulado.

**Tabla con el análisis de desviación para la variable altura**

Efecto	Desviación <sup>a</sup>	LRT <sup>b</sup> Chi-cuadrado	Componentes de varianza	Coefficientes de determinación
Familias	1428.81	0.00 NS	Va=19.65	h <sup>2</sup> a=0.004
Parcela	1428.81	0.00 NS	V <sub>parc</sub> =2043.50	C <sup>2</sup> <sub>parc</sub> =0.4628
LocalidadxFamilia	1447.40	18.59 **	V <sub>int</sub> =14.07	C <sup>2</sup> <sub>int</sub> =0.003
Modelo completo	1428.81		Ve=2338.26	

<sup>a</sup>Desviación del modelo ajustado sin los efectos correspondientes;

<sup>b</sup>LRT = Razón de verosimilitud;

\*\* = significativo por la prueba Chi-cuadrado con 1 grado de libertad, al 1% de probabilidad.

**6. Análisis de poblaciones o procedencias con progenies de medios hermanos dentro de la procedencia. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en una sola localidad.**

**Fichero: DatapinasterpoblaHS1loca-plantas.xls (Zas et al., 2004)**

El fichero con datos de *Pinus pinaster* tiene la siguiente estructura: Diseño en bloques completos al azar, 3 poblaciones, 3 familias/población, 2 repeticiones y 2 árboles/parcela.

VARIABLES medidas a los 8 años: variable 1 = altura (cm), variable 2 = diámetro a 1,3 m (cm) y variable 3 = volumen (diámetro<sup>2</sup> x altura/1000) en dm<sup>3</sup>.

Las columnas se configuran en formato número.

El orden de las columnas debe de ser el siguiente:

Individuo Familia Repetición Parcela Población Árbol variable1 variable2 variable3

**Objetivos: Determinar los componentes de la varianza y los valores genotípicos para la variable volumen.**

Una vez creado el fichero de datos en Excel y guardado en formato texto (MS-DOS) se guarda en la carpeta de Selegen y se hace clic en el fichero:

SELEGEN\_REMLBLUP\_W.exe

Clic en Si

Se abre una ventana del programa.

Selección de Várias populações ou Procedências

Arquivo

Selecciono el archivo: **DatapinasterpoblaHS1loca-plantas.txt**

## Modelo

Selecciono 005. Blocos ao acaso, progênies de meios-irmãos, várias plantas por parcela, várias populações (Bloques completos, progenies de medios hermanos, varias plantas por parcela, varias poblaciones)

$$y = Xr + Za + Wp + Ts + e$$

y = vector de datos fenotípicos

r = vector efecto repetición (fijo)

a = vector efectos genéticos aditivos individuales (aleatorio)

p = vector efecto parcela (aleatorio)

s = vector de los efectos de población o procedencia (aleatorios)

e = vector de error o residual (aleatorios)

X, Z, W y T son las matrices de incidencia para los efectos mencionados

Nº de variables: 3

Orden de variables a ser analizadas: 3.

Las demás opciones se dejan según aparecen en el modelo.

## Clic en Executar

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Actual : Dezembro 2020 - 28 ANOS - 200 Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Viçosa      Viçosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16: 330-339, 201

Arquivo      : C:\Users\Oliveira\Desktop\Selegen\Selegen_20190424_E\DataspinasterpoblaHS1loca-plantas.txt
Modelo       : 5
Numero de Variaveis : 3
Variavel Analisada : 3
N. Dados Lidos : 36
Zeros significativos : Nao
Deviance     : 190.36
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63           2.5% = 3.84
5% = 3.84           5% = 2.71
10% = 2.71
Data : 06/05/2024
Hora : 13:07:04
```

## Componentes de la varianza

### 1. Componentes de Variância ( REML Individual )

Va	=	7.624867
Vparc	=	1.074600
Vproc	=	0.011974
Ve	=	75.412822
Vf	=	84.124262
h2a	=	0.090638 +- 0.2838
c2parc	=	0.012774
c2proc	=	0.000142
CVgi%	=	13.289757
CVgp%	=	6.644878
CVe%	=	31.056894
Média geral	=	20.777778

### 2. Componentes de Média ( BLUP Individual )

#### Seleção de Procedencias

Ordem	Proc	g	u + g	Ganho	Nova Média
1	2	0.0030	20.7808	0.0030	20.7808
2	3	-0.0008	20.7769	0.0011	20.7789
3	1	-0.0021	20.7756	0.0000	20.7778

## Valores genotípicos

#### Seleção de Individuos

Ordem	Bloco	Familia	Proc	Árvore	f	a	u+a	Ganho	Nova Média	Ne	d	g
1	2	7	3	1	39.0000	1.8364	22.6142	1.8364	22.6142	1.0000	0.7511	2.5875
2	1	7	3	1	30.0000	1.3893	22.1671	1.6129	22.3906	1.6000	0.4530	1.8423
3	2	5	2	1	37.0000	1.2495	22.0273	1.4918	22.2695	2.4828	0.6790	1.9286
4	1	4	2	2	33.0000	1.1667	21.9445	1.4105	22.1883	3.4909	0.6238	1.7905
5	2	4	2	1	33.0000	0.9767	21.7545	1.3237	22.1015	4.1096	0.4971	1.4738
6	2	7	3	2	26.0000	0.9201	21.6978	1.2565	22.0342	4.3636	0.1402	1.0603
7	1	6	2	2	32.0000	0.8928	21.6706	1.2045	21.9823	5.3093	0.5809	1.4737
8	1	7	3	2	22.0000	0.8254	21.6032	1.1571	21.9349	5.3333	0.0771	0.9025
9	1	3	1	1	29.0000	0.7469	21.5247	1.1115	21.8893	6.2669	0.4309	1.1778
10	1	3	1	2	28.0000	0.6764	21.4542	1.0680	21.8458	6.9565	0.3839	1.0603
11	2	5	2	2	28.0000	0.6152	21.3929	1.0269	21.8046	7.6582	0.2561	0.8713
12	1	2	1	1	27.0000	0.5042	21.2820	0.9833	21.7611	8.5714	0.3530	0.8572
13	2	8	3	2	30.0000	0.4001	21.1778	0.9384	21.7162	9.5020	0.3806	0.7807
14	1	1	1	2	28.0000	0.2100	20.9878	0.8864	21.6642	10.4434	0.4225	0.6325
15	2	4	2	2	22.0000	0.2014	20.9791	0.8407	21.6185	10.8085	-0.0198	0.1816
16	2	3	1	1	21.0000	0.0312	20.8089	0.7901	21.5679	11.2000	-0.0463	-0.0151
17	1	5	2	2	16.0000	-0.0224	20.7553	0.7423	21.5201	11.6118	-0.1690	-0.1914

**7. Análisis de ensayos con clones no emparentados. Diseño en bloques completos al azar con varias plantas por parcela. Evaluación en varias localidades. Estabilidad, Adaptabilidad y Productividad.**

**Fichero: DatapinasterClones4loca-plantas.xls (Zas et al., 2004)**

El fichero con datos de *Pinus pinaster* tiene la siguiente estructura: Diseño en bloques completos al azar, 4 localidades, 10 clones, 2 repeticiones/localidad y 2 árboles/parcela.

VARIABLES medidas a los 8 años: variable 1 = altura (cm), variable 2 = diámetro a 1,3 m (cm) y variable 3 = volumen (diámetro<sup>2</sup> x altura/1000) en dm<sup>3</sup>.

Las columnas se configuran en formato número.

El orden de las columnas debe de ser el siguiente:

Localidad Individuo Genotipo Repetición Parcela Interacción Obs/parcela variable1  
variable2 variable3

**Objetivos: Determinar los componentes de la varianza, los valores genotípicos y los intervalos de confianza, la Estabilidad, Adaptabilidad y Productividad de la variable volumen.**

Una vez creado el fichero de datos en Excel y guardado en formato texto (MS-DOS) lo guardo en la carpeta de Selegen y hago clic en el fichero:

SELEGENREMLBLUP\_W.exe

Clic en Si

Se abre una ventana del programa.

Selección de Productivade, Estabilidade e Adapatabilidade

Archivo

Selecciono el archivo: **DatapinasterClones4loca-plantas.txt**

Delineamento Experimental

Blocos ao Acaso

## Modelo

Selecciono modelo 51: Estabilidade, Adaptabilidade e Produtividade: Blocos, genótipos, várias plantas por parcela, vários locais (Estabilidade, Adaptabilidade y Productividad: Bloques, genotipos, varias plantas por parcela, varias localidades)

$$y = Xr + Zg + Wp + Ti + e$$

y = vector de datos fenotípicos

r = vector efecto repetición (fijo)

g = vector efectos genotípicos (aleatorio)

p = vector efecto parcela (aleatorio)

i = vector de los efectos de interacción genotipo x ambiente (aleatorios)

e = vector de error o residual (aleatorios)

X, Z, W y T son las matrices de incidencia para los efectos mencionados

Nº de variables: 3

Orden de variables a ser analizadas: 3.

Las demás opciones se dejan según aparecen en el modelo.

Clic en Executar

## Resultados de la variable 3 (Volumen)

SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS - 200 Modelos  
Sistema Estatístico e Selecao Genética Computadorizada  
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária  
Embrapa Florestas  
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende  
Universidade Federal de Viçosa Viçosa - MG  
marcos.deon@gmail.com  
marcos.deon@ufv.br

Citação: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechnology 16:

```
Arquivo      : C:\Users\Usuario\Desktop\Selegen\Selegen_20190424_E\DatapinasterClones4loca-plantasMod51.txt
Modelo       : 51
Numero de Variaveis : 3
Variavel Analisada : 3
N. Dados Lidos : 160
Zeros significativos : Nao
Deviance     : 957.45
Valores Qui-Quadrado
1 grau de liberdade 0.5 grau de liberdade
1% = 6.63          2.5% = 3.84
5% = 3.84          5% = 2.71
10% = 2.71
Data : 04/12/2023
Hora : 13:41:48
```

## Componentes de la varianza

### 1. Componentes de Variância ( REML Individual )

Vg	=	2.816505
Vparc	=	108.743488
Vint	=	1.248346
Ve	=	96.759947
Vf	=	209.568286
h2g	=	0.013440 +- 0.0259
c2parc	=	0.518893
c2int	=	0.005957
h2mg	=	0.123699
Acgen	=	0.351709
rgloc	=	0.692893
CVgi%	=	6.780787
CVe%	=	50.646019
Média geral	=	24.750000

**Média geral = MG**

## Valores genotípicos

### 2. Componentes de Média ( BLUP Individual )

#### Seleção de Genotipos - Todos Locais

Ordem	Genotipo	g	u + g	Ganho	Nova Média	u+g+gem
1	5	0.7498	25.4998	0.7498	25.4998	25.5829
2	3	0.6416	25.3916	0.6957	25.4457	25.4627
3	4	0.5102	25.2602	0.6338	25.3838	25.3167
4	7	0.2319	24.9819	0.5334	25.2834	25.0076
5	1	0.1932	24.9432	0.4653	25.2153	24.9647
6	8	-0.0541	24.6959	0.3788	25.1288	24.6899
7	6	-0.1778	24.5722	0.2993	25.0493	24.5525
8	2	-0.3092	24.4408	0.2232	24.9732	24.4066
9	9	-0.6802	24.0698	0.1228	24.8728	23.9944
10	10	-1.1054	23.6446	0.0000	24.7500	23.5222

Local	Ordem	Genotipo	g+ge	u+g+ge	Ganho Nova	Media
1	1	3	0.7549	20.1299	0.7549	20.1299
	2	5	0.6891	20.0641	0.7220	20.0970
	3	4	0.4336	19.8086	0.6259	20.0009
	4	1	0.2705	19.6455	0.5370	19.9120
	5	7	0.2537	19.6287	0.4804	19.8554
	6	6	-0.1025	19.2725	0.3832	19.7582
	7	8	-0.1336	19.2414	0.3094	19.6844
	8	2	-0.3102	19.0648	0.2319	19.6069
	9	9	-0.7264	18.6486	0.1255	19.5005
	10	10	-1.1292	18.2458	0.0000	19.3750
2	1	5	1.0052	26.9552	1.0052	26.9552
	2	4	0.7732	26.7232	0.8892	26.8392
	3	3	0.7263	26.6763	0.8349	26.7849
	4	1	0.1753	26.1253	0.6700	26.6200
	5	7	0.1585	26.1085	0.5677	26.5177
	6	8	-0.1151	25.8349	0.4539	26.4039
	7	6	-0.1625	25.7875	0.3658	26.3158
	8	2	-0.3780	25.5720	0.2729	26.2229
	9	9	-0.8333	25.1167	0.1500	26.1000
	10	10	-1.3497	24.6003	0.0000	25.9500
3	1	5	0.8297	28.4797	0.8297	28.4797
	2	3	0.6997	28.3497	0.7647	28.4147
	3	4	0.5547	28.2047	0.6947	28.3447
	4	7	0.4296	28.0796	0.6284	28.2784
	5	8	0.0384	27.6884	0.5104	28.1604
	6	1	-0.0471	27.6029	0.4175	28.0675
	7	6	-0.1891	27.4609	0.3308	27.9808
	8	2	-0.4869	27.1631	0.2286	27.8786
	9	9	-0.6249	27.0251	0.1338	27.7838
	10	10	-1.2040	26.4460	0.0000	27.6500
4	1	5	0.8081	26.8331	0.8081	26.8331
	2	3	0.6703	26.6953	0.7392	26.7642
	3	4	0.5057	26.5307	0.6614	26.6864
	4	1	0.4601	26.4851	0.6111	26.6361
	5	7	0.1887	26.2137	0.5266	26.5516
	6	8	-0.0302	25.9948	0.4338	26.4588
	7	2	-0.1990	25.8260	0.3434	26.3684
	8	6	-0.3360	25.6890	0.2585	26.2835
	9	9	-0.8384	25.1866	0.1366	26.1616
	10	10	-1.2294	24.7956	0.0000	26.0250

Estabilidade de Valores Geneticos ( MHVG )

Ordem	Genotipo	MHVG
1	5	25.1069
2	3	25.0160
3	4	24.8368
4	7	24.5386
5	1	24.5151
6	8	24.2037
7	6	24.0952
8	2	23.9387
9	9	23.5096
10	10	23.0415

Adaptabilidade de Valores Geneticos ( PRVG )

Ordem	Genotipo	PRVG	PRVG*MG
1	5	1.0338	25.5875
2	3	1.0295	25.4802
3	4	1.0229	25.3172
4	7	1.0105	25.0098
5	1	1.0092	24.9770
6	8	0.9972	24.6813
7	6	0.9922	24.5563
8	2	0.9860	24.4045
9	9	0.9689	23.9802
10	10	0.9497	23.5059

Estabilidade e Adaptabilidade de Valores Geneticos ( MHPRVG )

Ordem	Genotipo	MHPRVG	MHPRVG*MG
1	5	1.0338	25.5872
2	3	1.0295	25.4795
3	4	1.0229	25.3168
4	7	1.0105	25.0094
5	1	1.0091	24.9757
6	8	0.9972	24.6811
7	6	0.9922	24.5561
8	2	0.9860	24.4042
9	9	0.9689	23.9794
10	10	0.9497	23.5051

## 8. Índice de selección

Una vez corridas las tres variables con el modelo 1 o con el modelo 4 (varias localidades), minimizar la pantalla principal de SELEGEN y se selecciona “Índice de Selecao”

Se deja la opción “Aditivo” que es el índice de selección Smith-Hazel

Clic en Executar

Aparece:

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS ò 200
Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Vicosa      Vicosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citacao: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechn
ology 16: 330-339, 2016.
Direcao ( maior, menor, nula ) e Peso da Variavel  1 :
```

Introducir los pesos económicos del Índice de Selección

Para la variable 1 un peso de 0.6

Maior0.6 Enter

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS ò 200
Modelos
Sistema Estatístico e Selecao Genética Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
               Universidade Federal de Vicosa      Vicosa - MG
               marcos.deon@gmail.com
               marcos.deon@ufv.br

Citacao: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechn
ology 16: 330-339, 2016.
Direcao ( maior, menor, nula ) e Peso da Variavel  1 :
maior0.6
Direcao ( maior, menor, nula ) e Peso da Variavel  2 :
```

Para la variable 2 un peso económico de 0.2

## Maior0.2 Enter

```
SELEGEN-REML/BLUP Inicio : 1993 Versao Atual : Dezembro 2020 - 28 ANOS Ò 200
Modelos
Sistema Estatístico e Selecao Genetica Computadorizada
Modelos Lineares Mistos via REML/BLUP e REML/GLS

Empresa Brasileira de Pesquisa Agropecuaria
Embrapa Florestas
Ministerio da Agricultura e do Abastecimento

Informacoes : Marcos Deon Vilela de Resende
Universidade Federal de Vicosa Vicosa - MG
marcos.deon@gmail.com
marcos.deon@ufv.br

Citacao: Resende, M.D.V. Software Selegen-REML/BLUP: a useful tool for plant breeding.Crop Breeding and Applied Biotechn
ology 16: 330-339, 2016.
Direcao ( maior, menor, nula ) e Peso da Variavel 1 :
maior0.6
Direcao ( maior, menor, nula ) e Peso da Variavel 2 :
maior0.2
Direcao ( maior, menor, nula ) e Peso da Variavel 3 :
```

Para la variable 3 un peso económico de 0.2

## Maior0.2 Enter

Ejemplo con el modelo 1.

### Valores Geneticos

Genotipo	Var. 1	Var. 2	Var. 3
1	445.6212	7.0004	25.1945
2	437.5224	6.9208	24.0662
3	425.8693	7.1066	23.9286
4	427.4716	7.6641	26.6256
5	387.7057	6.4960	20.8188
6	436.5319	7.7172	29.4051
7	471.4036	8.2482	34.7166
8	513.8789	8.4605	39.1473
9	389.0166	6.2305	16.4981
10	442.6788	6.5756	21.2591
Media	437.7700	7.2420	26.1660
Variancia	1349.1253	0.5672	45.4253
Desvio	36.7304	0.7531	6.7398

### Selecao de Genotipos

Ordem	Genitor	Indice	Ganho	Ganho %
1	8	11.8028	11.8028	19.8164
2	7	10.9211	11.3619	15.3411
3	6	10.0528	10.9255	10.9112
4	1	9.8860	10.6657	8.2729
5	4	9.8082	10.4942	6.5321
6	2	9.6991	10.3617	5.1868
7	10	9.6083	10.2540	4.0943
8	3	9.5540	10.1665	3.2060
9	5	8.6761	10.0009	1.5249
10	9	8.4988	9.8507	0.0000

Da un ranking genético para cada una de las tres variables, luego unos cuantos parámetros: media, varianza y desviación.

Finalmente aparece un índice para cada individuo con el valor mixto de ambas variables, la ganancia genética en esas mismas unidades y en porcentaje.

## 9. Agradecimientos

Al Dr. Olman Murillo Gamboa (Instituto Tecnológico de Costa Rica (TEC)) y al Dr. Marcos Deon Vilela de Resende (EMBRAPA de Brasil) por los consejos en la utilización del programa SELEGEN-REML/BLUP.

## 10. Bibliografía

Annicchiarico, P., 1992. Cultivar adaptation and recommendation from alfalfa trials in northern Italy. *Journal of Genetics and Plant Breeding*, 46: 269-278.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31: 423-447.

Lin, C.S., Binns, M.R., 1988. A superiority measure of cultivar performance for cultivar x location data. *Canadian Journal of Plant Science*, 68: 193-198.

Mendes, F.F., Guimarães, L.J.M., Souza, J.C., Guimarães, P.E.O., Pacheco, C.A.P., Machado, J.R.A., Meirelles, W.F., da Silva, A.R., Parentoni, S.N., 2012. Adaptability and stability of maize varieties using mixed model methodology. *Crop Breeding and Applied Biotechnology*, 12: 111-117.

McCullagh, P., Nelder, J., 1989. *Generalized Linear Models* (2 ed.). London: Chapman & Hall. London. 532 p.

Oliveira, J.A., Bande, M.J., 2024. Productividad, estabilidad y adaptabilidad de cultivares de maíz forrajero en Galicia. *Revista Vaca Pinta*, 41: 90-96.

Piepho, H.P., Möhring, J., Melchinger, A.E., Büchse, A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161: 209-228.

Pinto, Jr J.E., Sturion, J.A., Resende, M.D.V., Ronzelli, Jr P., 2006. Avaliação simultânea de produtividade, adaptabilidade e estabilidade genotípica de *Eucalyptus grandis* em distintos ambientes do Estado de São Paulo. *Boletim de Pesquisa Florestal* 53: 79-108.

Resende, M.D.V., 2007. Software SELEGEN-REML/BLUP: Sistema estatístico e seleção genética computarizada via modelos lineares mistos. Embrapa Florestas, Colombo, Brasil, 350 p.

Resende, M.D.V., 2016. Software Selegen-REML/BLUP: A useful tool for plant breeding. *Crop Breeding and Applied Biotechnology*, 9: 313-319.

Resende, M.D.V., Murillo, O., Badilla, Y., 2018. *Genética Cuantitativa y Selección en el Mejoramiento Forestal*. Editorial Tecnológica de Costa Rica. Cartago, Costa Rica. 302 p.

Verardi, C.K., Resende, M.D.V., Costa, R.B., Gonçalves, P.S., 2009. Adaptabilidade e estabilidade da produção de borracha e seleção em progênies de seringueira. *Pesquisa Agropecuária Brasileira*, 44: 1277-1282.

Zas, R., Merlo, E., Fernández-López, J., 2004. Genetic parameter estimates for maritime pine in the Atlantic Coast of northwest Spain. *Forest Genetics*, 11, 1: 45-53

Zobel, B., Talbert, J., 1988. *Técnicas de mejoramiento genético de árboles forestales*. Editorial Limusa. México. 545 p.