Rutinas estadísticas para el programa MURRAP

Translation Prepared by:
Rita R. Plá
rpla@cae.cnea.gov.ar
Técnicas Analíticas Nucleares – CAE
Comisión Nacional de Energía Atómica
Av. Del Libertador 8250
Buenos Aires 1429 Argentina

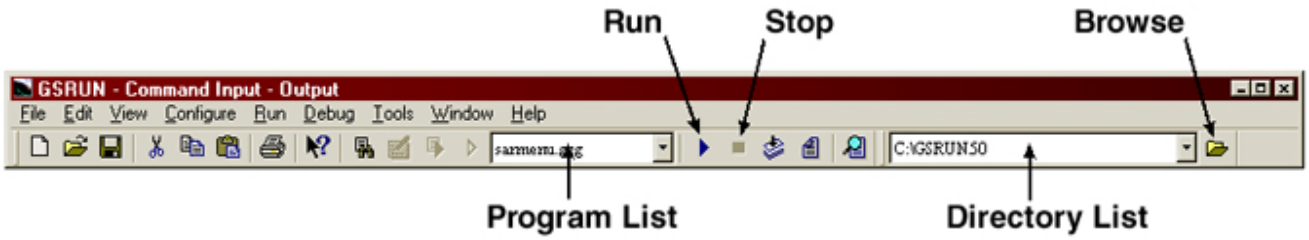
Información General

Estos programas se basan en las rutinas SARCAR (Smithsonian Archaeometric Research Collections and Records) y fueron traducidos al lenguaje Gauss por Hector Neff. Las revisiones para compatibilizar estas rutinas con la Gauss-Run time GUI y programación adicional (incluyendo la redacción de estas páginas de ayuda) fueron realizadas por Bill Grimm.

La Run-Gauss time GUI (interfaz gráfica de usuario) es un software libre producido por Aptech, Inc. Se ha incluido un archivo de ayuda en la interfaz gráfica de usuario, con instrucciones sobre su uso. Las siguientes instrucciones contienen aspectos particulares de la operación de la interfaz gráfica de usuario.

Cómo empezar

Usted tiene que haber recibido este programa de Mike Glascock del MURR (o de alguno de sus asociados). Por lo tanto, ya deber tener instalado el Gauss Run-time y descomprimidas las rutinas MURR. Las cosas importantes que deben conocerse acerca del Gauss GUI están en la barra de herramientas:



El ítem más útil es la lista de Directorios (Directory List). El directorio (donde se encuentran ubicados sus archivos de datos) y esta lista, que conserva los directorios utilizado más recientemente, pueden cambiarse utilizando el botón de navegación. El botón Detener (Stop) detendrá una rutina, pero si está esperando un input, será necesario pulsar la tecla Intro (Enter) o Retorno (Return) para realmente detener el programa. Para que el Listado de Programas (Program List) muestre algo (la primera vez que ejecute el programa estará en blanco), se debe salir del programa ingresando la opción 16 y luego escribir run sarmenu.gcg y pulsar la tecla Enter. Ahora debera aparecer sarmenu.gcg en la Program List, y si detiene el programa (o si se detiene porque ha detectado un error) se puede reiniciar el menú pulsando el botón Ejecutar (Run).

Usando el Menú

Las rutinas MURR comienzan con una página de menú (sarmenu.gcg) con la lista de 16 opciones disponibles. Para seleccionar una opción, introduzca el número de esa opción en el prompt. A continuación, deberá ingresar la información requerida por cada rutina. Para una descripción de las rutinas disponibles, haga clic en Opciones de Menú de Programa (Program Menu Options) a la izquierda y seleccione la opción de la que desee obtener más información.

Una cosa para recordar siempre que utilice estas rutinas, es que nunca deberá incluir la extensión del archivo cuando introduzca el nombre de un archivo o conjunto de datos (salvo en el caso de especificar un archivo de salida, donde probablemente querrá añadir la extensión .txt).

Archivos de datos

Los archivos de datos utilizados por este programa están en formato propietario Gauss. Se ha incluido una rutina para importar datos desde Microsoft Excel. La mayoría de los archivos importados de Excel deben ser formateados de la siguiente manera (algunas

rutinas utilizan archivos con configuraciones ligeramente diferentes y esas diferencias figuran en la página de ayuda para esas rutinas):

	A	B	C	D	E	F
1	Anid	La	Lu	Nd	Sm	Yb
2	BG046	37.38	0.4703	29.681	6.45	2.964
3	BG047	37.161	0.3884	34.71	6.145	2.886
4	BG048	22.864	0.2592	24.038	3.942	1.991
5	BG049	38.397	0.3793	30.007	6.427	3.045
6	BG050	36.566	0.4635	30.786	6.34	3.12
7	BG051	36.764	0.4455	29.468	6.29	3.157

La primera fila debe contener el nombre del encabezamiento para la columna de las muestras (Anid) y los nombres de las variables utilizadas (As, La, Lu, etc.). Los nombres asignados a las columnas no pueden tener más de ocho caracteres. A partir de la segunda fila, las columnas deben contener los siguientes datos:

Columna 1: columna de texto que contiene los nombres de las muestras (no debe exceder los ocho caracteres alfanuméricos).
Columnas 2, 3, etc: columnas numéricas que contienen datos de concentración elemental (sólo pueden contener números).

Si el archivo de Excel fue editado (eliminando filas o columnas), hay que asegurarse de guardarlo y siempre cerrar Excel antes de de importar el archivo a Gauss. De lo contrario, podrían producirse errores aparentemente inexplicables.

Un comentario sobre Excel:

Si alguna vez le ocurre que no puede abrir Excel o si obtiene errores muy peculiares al intentar abrir un archivo con Excel haciendo doble clic en él, busque en su pc un archivo llamado Excel.xlb. Seguramente se encontrará en la carpeta C: \ Windows \ Application Data \ Microsoft \ Excel. Cuando lo encuentre, elimínelo. Excel hará uno nuevo cuando vuelva a comenzar y ésto por lo general, resuelve cualquier problema de comienzo.

GAUSS

Opciones del menú

Here are the options available: (Opciones disponibles)

1. Specimen sourcing routines.
2. Missing values substitution.
3. Transform a dataset.
4. Dataset manipulation routines.
5. List values of variables in a dataset.
6. Obtain summary statistics for a dataset.
7. Hierarchical cluster analysis.
8. Principal components analysis.
9. Canonical discriminant analysis.
10. Classify specimens using Mahalanobis distance.
11. Project specimens using Mahalanobis distance.
12. 2D scatterplots.
13. 3D scatterplots.
14. Convert all v89 files to v96 format in current directory.
15. Help files.
16. Exit program.

Choice ?

1 – Specimen sourcing routines (Rutinas de "sourcing" de casos)

Estas rutinas usan diversos métodos para determinar cuáles muestras de referencia se

encuentran cerca de otras pertenecientes al conjunto de datos desconocidos. Al seleccionar esta opción, encontrará otras tres opciones:

1. Mean euclidean distance searches. (Búsquedas por distancia euclideana media)
2. Sourcing for specimens with x-y coordinates. ("Sourcing" de casos con coordenadas x-y)
3. Sourcing for specimens without x-y coordinates. ("Sourcing" de casos sin coordenadas x-y)

Choice? (Elección)

La primera opción utiliza la distancia euclideana media para comparar muestras de dos grupos. Las otras dos utilizan la distancia de Mahalanobis para comparar muestras desconocidas con un grupo conocido de muestras (la primera de estas opciones también toma en cuenta datos geográficos o UTM).

Mean Euclidean distance searches (Búsquedas por distancia euclideana media)

Búsqueda por distancia euclideana media de un conjunto de datos log-transformado.

Asegúrese de que el conjunto de datos de referencia tenga la misma estructura que el conjunto de datos que contiene los casos desconocidos.

Esta rutina asume que se tienen al menos diez muestras en el conjunto de datos de referencia.

La estructura (número y orden de las columnas) de los dos conjuntos de datos a comparar tiene que ser la misma. No existen rutinas de chequeo de errores para verificar esto, de manera que usted mismo debe asegurarse de que las estructuras sean iguales. Una manera de hacer esto es usar la opción "View/edit a dataset using Excel" (Visualizar/editar un conjunto de datos con Excel) que se encuentra en la opción #4 del menú principal (rutinas de manipulación de conjuntos de datos -Dataset Manipulation routines).

También es necesario que el conjunto de datos de referencia contenga al menos diez muestras o se obtendrá un error al correr esta rutina. El objetivo de esta rutina es dar un listado de las diez muestras del conjunto de datos de referencia más cercanas a cada una de las muestras en el conjunto de datos desconocidos, empleando la distancia euclideana media como indicador de disimilitud (números más grandes = más disímiles). Se usa la distancia euclideana media debido a que es menos sensible a la existencia de datos faltantes (si los hubiera). La rutina también supone que los dos conjuntos de datos están log-transformados para reducir al mínimo las diferencias de escala entre las variables.

Enter name of reference dataset to search*:*

Ingrese el nombre del conjunto de datos de referencia (que tiene al menos diez muestras), sin incluir la extensión del archivo (*mydata* en lugar de mydata*.dat*).

Do you wish to select a subset of variables? ¿Desea seleccionar un subconjunto de variables? (por defecto = N)

Si no desea usar todas las variables disponibles para calcular las distancias, introduzca Y. Si decide seleccionar un subconjunto de variables, se verá algo como lo siguiente:

Here are the variables in your dataset: (Aquí están las variables de su conjunto de datos):

Name Index
LA 1
LU 2
ND 3

SM 4
U 5
YB 6
CE 7
CO 8
CR 9
CS 10
EU 11
FE 12
HF 13
RB 14
SB 15
SC 16
SR 17
TA 18
TB 19
TH 20
ZN 21
ZR 22
AL 23
BA 24
CA 25
DY 26
K 27
MN 28
NA 29
TI 30
V 31

Index of variable to keep (0 to end selection)? (Índice de variable a mantener (0 para poner fin a la selección)?

Seleccione las variables que desea incluir en la medición de distancia ingresando el número de índice asociados a ellas. Cuando haya terminado, ingrese 0.

Enter name of unknown specimens dataset: (Introduzca el nombre del conjunto de datos desconocido)

Este es el nombre del archivo de datos que desea comparar con el conjunto de datos de referencia. Una vez más, no se incluye la extensión al ingresar el nombre.

Ouput file name: (Nombre del archivo de salida)

La rutina crea un archivo de texto con los resultados de los cálculos. En este caso, usted querrá incluir la extensión (i.e. myoutput.txt). Este es un ejemplo de los resultados a esperar:

Here are the 10 closest specimens to BG001 (10 casos más cercanos a BG001)
BG001 0.000000
BG186 0.352963
BG291 0.493282
BG174 0.536364
BG171 0.817943
BG271 0.915002
BG138 0.929019
BG002 0.988339
BG179 1.162293
BG214 1.179396

Here are the 10 closest specimens to BG002 (10 casos más cercanos a BG002)
BG002 0.000000
BG291 0.797076
BG001 0.988339

BG174 1.160463
BG186 1.322073
BG138 1.548494
BG299 1.567799
BG146 1.707235
BG171 1.792749
BG236 1.826634

Here are the 10 closest specimens to BG003 (10 casos más cercanos a BG003)
BG003 0.000000
BG007 0.800994
BG056 0.986216
BG066 1.272771
BG082 1.316656
BG064 1.477696
BG085 1.573144
BG094 1.651708
BG069 1.701913
BG074 1.772872

Sourcing for specimens with x-y coordinates ("Sourcing" de muestras con coordenadas xy)

Calculo de distancia de Mahalanobis para comparar los distintos casos individuales en dos o más grupos sobre la base de la matriz de varianza-covarianza de uno de los grupos (el 'conocido')

Nota:
KNOWN group (grupo conocido) las tres primeras columnas deben estar encabezadas por anid, x, e y en;
UNKNOWN group (grupo desconocido) debe tener variables en el mismo orden, pero sin x e y en las columnas 2-3; por lo tanto, el UNKNOWN group tiene 2 columnas menos que el KNOWN group.

KNOWN group debe tener al menos dos casos más que variables. Se le ofrecerá la oportunidad de seleccionar un subconjunto de variables.

Los valores faltantes en el KNOWn group serán reemplazados con los valores del mejor ajuste (best fit values). Los valores faltantes en los conjuntos de datos de comparación se calcularán para minimizar la distancia de Mahalanobis desde cada grupo de referencia con el que se quieran hacer las comparaciones.

Esta rutina utiliza la distancia de Mahalanobis junto con datos de proximidad geográfica, para determinar la probabilidad de que los casos desconocidos individuales estén en el grupo conocido (KNOWN group). Probablemente deberá transformar los datos logaritmicamente para reducir las diferencias de escala entre las variables. El KNOWN group también necesita tener información geográfica (UTMs, datos de procedencia o algún tipo de datos x-y) en la segunda y tercera columnas del conjunto de datos. Si está utilizando UTMs, puede que también desee transformar estas medidas.

Enter name for KNOWN group (assumed to have x in 2, y in 3) (Introduzca el nombre del grupo Conocido (asumiendo tener x en 2, y en 3):

Introduzca el nombre del conjunto de datos conocidos o de referencia, sin incluir la extensión del archivo. La segunda y tercera columnas deben tener coordenadas geográficas de algún tipo. Este conjunto de datos también tiene que tener por lo menos dos muestras más que el número de variables que contiene (incluyendo los datos de las coordenadas x-y).

Do you wish to select a subset of variables (default=N)? (¿Desea seleccionar un subconjunto de variables (por defecto = N)?)

Si no desea usar todas las variables disponibles para calcular las probabilidades, ingrese Y. Si decide seleccionar un subconjunto de variables, verá algo como lo siguiente:

Here are the variables in your dataset (Variables en su conjunto de datos):

Name	*Index*
LA	*1*
LU	*2*
ND	*3*
SM	*4*
U	*5*
YB	*6*
CE	*7*
CO	*8*
CR	*9*
CS	*10*
EU	*11*
FE	*12*
HF	*13*
RB	*14*
SB	*15*
SC	*16*
SR	*17*
TA	*18*
TB	*19*
TH	*20*
ZN	*21*
ZR	*22*
AL	*23*
BA	*24*
CA	*25*
DY	*26*
K	*27*
MN	*28*
NA	*29*
TI	*30*
V	*31*

Index of variable to keep (0 to end selection) (Índice de variable a mantener (0 para poner fin a la selección)?

Seleccione cuales variables desea incluir en los cálculos de probabilidad entrando el número de índice asociado con cada una de ellas. Cuando haya terminado, introduzca 0.

Enter name for UNKNOWN specimen dataset (*Ingrese el nombre del conjunto de datos de los casos desconocidos):*

Este es el nombre del archivo de datos que desea comparar con el conjunto de datos de referencia. Una vez más, no se incluye la extensión cuando introduzca el nombre.

Enter name for dataset for variante-covariance matriz (must be structured exactly like known and unknown data) Introduzca el nombre del archivo de datos para la matriz de varianza-covarianza (Debe tener exactamente una estructura igual que los datos conocidos y desconocidos):

A menos que tenga una buena razón para hacerlo de otra manera, debe ingresar el nombre del conjunto de datos conocidos (sin incluir la extensión del archivo).

Output file name: (Nombre del archivo de salida)

La rutina crea un archivo de texto con los resultados de los cálculos. En este caso, usted deberá incluir la extensión. La salida consiste en una matriz donde las filas son los casos del grupo de referencia y las columnas son los casos desconocidos.

Sourcing for specimens without x-y coordinates. ("Sourcing" de casos sin coordenadas x-y

Esto es esencialmente igual que la rutina para los casos con coordenadas x-y, con la excepción de que el grupo conocido contiene todas las mismas variables que el desconocido y que la matriz de varianza-covarianza (la segunda y tercera columnas no son datos x-y).

2. Missing values substitution.(Sustitución de valores faltantes)

Esta rutina utiliza la distancia de Mahalanobis para sustituir los valores faltantes en el conjunto de datos con aproximaciones. Intenta minimizar la distancia de Mahalanobis entre la muestra con los datos faltantes y el centroide del conjunto de datos. No utilice esta rutina si tiene más de un grupo representado en el conjunto de datos, o de lo contrario las aproximaciones serán sesgadas. También deberá tener por lo menos dos casos más que variables para que la rutina corra. El conjunto de datos que abra con esta rutina será sustituido con aquel con los valores faltantes reemplazados, por lo que le recomendamos hacer una copia de la base de datos antes de ejecutar la rutina.

Al igual que con todas las rutinas, no incluya la extensión cuando ingrese el nombre del archivo de su conjunto de datos.

3. Transform a dataset (Transformar un conjunto de datos)

Esta rutina le permite transformar los datos en un conjunto de datos Gauss y luego guardar el nuevo archivo con otro nombre. No es necesario incluir la extensión para los archivos de entrada o salida, ya que se entiende que ambos son conjuntos de datos Gauss. Las opciones de transformación son las siguientes:

1. Transform raw ppm to log base 10 of ppm
2. Transform raw ppm to natural logs
3. Tranform log base 10 data to ppm
4. Standardize ppm data
5. Standardize against another dataset
6. Temper a raw clay dataset
7. Best relative fit for logged data to group mean
8. Best relative fit for logged data to mean of another group
9. Centered log ratio for log-10 data (Aitchison's transformation)

Transformation?

1 – Transform raw ppm to log base 10 of ppm (Transformar ppm a log ppm en base 10)

Esta opción le permite transformar una parte o la totalidad de los datos en un conjunto de datos Gauss, de partes por millón (ppm o mg / kg) a sus valores de logaritmos en base 10. Esta transformación ayuda a igualar las escalas de los diferentes elementos de forma que aquellos con una alta concentración (como el hierro o el aluminio) no eclipsen las concentraciones de las tierras raras. Basta con introducir el nombre del archivo de datos que desea transformar (sin la extensión del archivo), seleccionar la opción # 1, especificar un nuevo nombre para los datos transformados (sin la extensión) y, a continuación, determinar, si es su caso, las variables que desea omitir en la transformación. Si hubieran datos faltantes, se le preguntará si desea reemplazarlos de inmediato (tiene la opción de hacerlo por separado a través del menú principal con la opción # 3). De cualquier manera que lo haga, los valores faltantes se insertarán en el nuevo archivo (datos transformados) y no en la base de datos original. Si desea tener una copia de los valores transformados sin los valores faltantes sustituidos, puede optar por no sustituirlos como parte de esta rutina, salir y guardar una copia del archivo con otro nombre y luego seleccionar la opción de sustitución de valores faltantes (Missing values substitution option) en el menú principal,

para proceder a la sustitución de los valores. Esto sirve para todas las siguientes transformaciones que brindan la opción de completar los valores faltantes (todas las relacionadas con transformaciones logarítmicas).

2 – Transform raw ppm to natural logs (Transformar ppm a logaritmos naturales de ppm)

Esta opción le permite transformar una parte o la totalidad de los datos en un conjunto de datos Gauss, de partes por millón (ppm o mg / kg) a los valores de logaritmos naturales (base e). Véase más arriba para más detalles.

3 - Transform log base 10 data to ppm (Transformar datos log base 10 a ppm)

Esta opción le permite transformar una parte o la totalidad de los datos de un conjunto de datos Gauss de logaritmos en base 10 a partes por millón (ppm o mg / kg). Véase más arriba para más detalles.

4 - Standardize ppm data (Estandarizar datos en ppm)

Esta opción le permite estandarizar algunos o todos los datos en un conjunto de datos Gauss. Véase más arriba para más detalles.

5 - Standardize against another dataset (Estandarizar contra otro conjunto de datos)

Esta opción le permite estandarizar algunos o todos los datos en un conjunto de datos Gauss utilizando la media y desvío estándar de otro conjunto de datos. Véase más arriba para más detalles.

6 - Temper a raw clay dataset (Agregar antiplástico a un conjunto de datos de arcilla cruda)

Esta opción le permite simular matemáticamente el agregado de antiplástico a un conjunto de datos de arcilla. Las muestras de antiplástico y de arcilla deberán estar en conjuntos de datos separados y ambos deberán ser transformados logarítmicamente antes de usar esta opción. Se le pedirá ingresar un valor de concentración de antiplástico (ingresar un número entre 0 y 1). Para cada muestra de arcilla en su base de datos, la rutina aleatoriamente seleccionará una muestra de antiplástico del conjunto y la combinará con esa muestra de arcilla. Asegúrese de tener las mismas variables, en el mismo orden, en ambos conjuntos de datos antes de usar este procedimiento.

7 - Best relative fit for logged data to group mean (Mejor ajuste relativo de datos log transformados a la media grupal)

Esta rutina se utiliza para corregir por correlaciones positivas de las concentraciones elementales que resulten de modificaciones de la pasta cerámica de origen natural o humano. Algunos ejemplos de esto podrían ser las inclusiones naturales de arena en la arcilla (la arena, compuesta mayoritariamente por sílice, no contribuye a las concentraciones elementales determinadas por AAN) o arena u otros materiales añadidos como antiplásticos por el alfarero. El mejor factor de ajuste relativo también se conoce como factor de dilución. Se debe tener un solo grupo en la base de datos o esta rutina no dará buenos resultados.

8 - Best relative fit for logged data to mean of another group (Mejor ajuste relativo de datos log transformados a la media de otro grupo)

Utilice esta rutina para encontrar el mejor ajuste relativo de un grupo desconocido a la media de un grupo conocido. Asegúrese de que su archivo de datos represente un solo grupo y que las variables para los dos conjuntos de datos sean las mismas y en el mismo orden.

9 - Centered log ratio for log-10 data (Aitchison's transformation)

Esta rutina calcula lo que se conoce en literatura como la transformación de Aitchison. Antes de ejecutarla, sustituir todos los valores faltantes en la base de datos y transformar los datos a logaritmo en base 10.

4. Dataset manipulation routines (Rutinas de manipulación de datos)

Al seleccionar esta opción, aparecen otras cinco opciones para elegir:

DATA MANIPULATION ROUTINES

1. *Start the Dataset Manipulator program.*
2. *View/edit a dataset using Excel.*
3. *Import data from Excel file.*
4. *Export data to Excel file.*
5. *Concatenate datasets.*

Choice?

1 – Start the dataset manipulator program (Inicie el programa manipulador de datos)

El programa Manipulador de datos fue escrito para facilitar el movimiento de muestras de un conjunto de datos a otro, o dividir un conjunto de datos en varios grupos para realizar un análisis posterior o gráficos. Funciona muy bien, pero debe ser cuidadoso y siempre guardar los archivos al terminar de añadir o eliminar muestras, antes de abrir un nuevo archivo para trabajar, o puede perder los cambios realizados a uno de los archivos. El programa incluye un archivo de ayuda muy sencilla para comenzar, pero es más bien un programa para aprender sobre la marcha. Lo hallará muy versátil una vez que se acostumbre a la forma en que funciona, y recuerde siempre guardar los archivos a medida que trabaja.

Tenga en cuenta también que el programa Dataset manipulador, como la mayoría de las rutinas Gauss, no verifica si sus columnas corresponden a las mismas variables. Usted es el que debe saber lo que hay en cada uno de sus archivos de datos.

2 – View/edit a dataset using Excel (Visualizar/editar un conjunto de datos usando Excel)

Close Excel to return to this program when finished (*Cerrar Excel para regresar a este programa cuando haya terminado).*

Name of file to view/edit (*Nombre de archivo para ver/editar):*

Escriba el nombre del archivo de datos que desea ver en Excel (no incluya la extensión del archivo).

Now Start Excel and open TEMPFILE in the current directory. If edits were made, you will need to save the file before closing Excel. When finished, press Enter. (Inicie Excel y abra TEMPFILE en el directorio actual. Si se hicieron modificaciones, tendrá que guardar el archivo antes de cerrar Excel. Cuando termine, pulse Intro).

Utilizando el Explorador (Explorer) (herramienta de Windows), vaya a su directorio de trabajo actual (mostrado en la lista de Directorios en la barra de herramientas de Gauss) y haga doble clic en *tempfile.xls*. Esto debería abrir el archivo en Excel, donde se puede ver el archivo o editarlo. Si se edita el archivo, deberá guardarlo antes de cerrar Excel y volver a Gauss. Al volver a Gauss, pulse Intro y se le dará la opción de actualizar el archivo. No es necesario actualizarlo si no hace ninguna modificación.

3 - Import data from Excel file (Importar datos de un archivo de Excel)

Esta rutina le permite convertir un archivo de Excel en un archivo de datos Gauss. Consulte la sección de archivos de datos de la Introducción de este documento para saber como debería ser la estructura de los datos (algunas rutinas requieren estructuras levemente diferentes, de modo que refiérase a la rutina específica que planea usar para cualquier excepción a estas reglas).

4 – Export data to Excel file (Exportar datos a un archivo de Excel)

Esta rutina le permite convertir un archivo de datos Gauss en un archivo de Excel con la misma estructura.

5 – Concatenate datasets (Concatenar conjuntos de datos)

Esta rutina proporciona una manera fácil de unir varios conjuntos de datos Gauss **con la misma estructura** con rapidez. Esta operación también puede ser realizada usando Excel o el Dataset Manipulador, pero esta rutina la hace mucho más simple. Asegúrese de que tiene la misma estructura en todos los conjuntos de datos que desea añadir.

5. List values of variables in a dataset (Listar valores de las variables en un conjunto de datos)

Utilice esta rutina si sólo desea ver rápidamente en la pantalla los valores de un conjunto de datos Gauss. Esta rutina no es adecuada para manipular datos, para eso debe utilizarse la opción 4.

Al iniciar la rutina, se le pedirá el nombre del archivo de datos del cual desea la lista (no incluya la extensión *.dat* al escribir el nombre). La lista es enviada sólo a la pantalla, pero si lo desea, se puede copiar y pegar en otro documento.

6. Obtain summary statistics for a dataset (Obtener un resumen estadístico para un conjunto de datos)

Esta rutina crea un archivo de texto con un listado de la estadística descriptiva para un conjunto de datos. Las estadísticas se calculan en unidades de concentraciones elementales (ppm). Tiene la opción de calcular, ya sea la media aritmética o geométrica y las desviaciones estándar para sus archivos de datos.

Al ejecutar el procedimiento, se le pide que seleccione el tipo de estadística que desea, aritmética o geométrica. A continuación, deberá ingresar un nombre para el archivo de salida (no olvide incluir la extensión. *.txt*). Si introduce un nombre de archivo en blanco, la rutina lo devolverá al menú principal. El programa entonces preguntará por el nombre del conjunto de datos Gauss del cual desea obtener el resumen, así como si los datos en el

archivo están o no transformados logarítmicamente. Por defecto, esto significa logaritmo base-10 y no logaritmos naturales. El listado de la estadística descriptiva se imprime, y se le pregunta si desea ejecutar lo mismo con otro conjunto de datos. Si decide no hacerlo, el programa lo devolverá al menú principal. El archivo acabado de crear puede encontrarse en el directorio de trabajo actual.

7. Hierarchical Cluster Análisis (Análisis de conglomerados jerárquico)

Hemos proporcionado una rutina simple de agrupación con el fin de ver rápidamente si hay algunos grupos evidentes en un dado conjunto de datos. Esta rutina calcula la distancia euclideana media entre todas las muestras y construye un dendrograma utilizando el algoritmo de vinculación media. La rutina no sustituye automáticamente los datos faltantes, así que asegúrese de hacerlo antes de ejecutarla o sus resultados pueden no ser satisfactorios.

Al ejecutar la rutina, primero se le pedirá que introduzca el nombre de su archivo de datos (no incluya la extensión .dat). Luego se le ofrecerá la opción de seleccionar un subconjunto de las variables presentes en su base de datos. Presione Enter para aceptar la opción por defecto (N), o pulse Y para seleccionar las variables. Cuando haya terminado de seleccionar las variables (si decide hacerlo), se le preguntará si el conjunto de datos ha sido transformado a logaritmo (base-10). Una vez que responda a esta pregunta, se le dará una posición calculada por defecto para las etiquetas de las muestras en el dendograma. Siempre acepte el valor por defecto la primera vez (por lo general es una buena estimación). El dendrograma se visualiza en la pantalla en una nueva ventana. Entonces, se le da la opción de cambiar la posición de la etiqueta.

Esta rutina tiene algunos defectos, pero es para ser utilizada principalmente como una herramienta de exploración - no como un medio para generar un dendrograma de calidad para publicar.

8. Principal Components Análisis (Análisis de componentes principales)

Esta rutina realiza un análisis de componentes principales (PCA) sobre un conjunto de datos Gauss. También puede proyectar un conjunto de datos en el espacio de los componentes principales de otro conjunto de datos. Esto es útil para graficar un conjunto de datos vs. otro, sobre ejes de componentes principales comunes. La rutina cuenta con muchas opciones para elegir, por lo que las mismas se discutirán individualmente.

Principal Components Análisis (*Análisis de componentes principales)*

Calculates principal components for a dataset and, optionally, projects other datasets into the principal components space (Calcula componentes principales para un conjunto de datos y, opcionalmente, proyecta otro conjunto de datos dentro del espacio de componentes principales.

Datasets must contain same varialbles , in same order (Los conjuntos de datos deben contener las mismas variables, en el mismo orden)

You should already have taken care of missing values in your data (Ya se debería haber ocupado de los datos faltantes).

Si desea proyectar un conjunto de datos en el espacio de los componentes principales de otro conjunto de datos, ambos deben tener las mismas variables en el mismo orden. Además, esta rutina no funcionará si tiene datos faltantes, así que asegúrese de solucionarlo antes de ejecutarla.

Here are your choices(Opciones)

1. Principal components analysis using variance-covariance matrix.(PCA usando matriz de varianza-covarianza)

2. Principal components analysis using correlation matrix.(PCA usando matriz de correlación)

3. Simultaneous RQ factor analysis with variance-covariance matrix.(FA RQ simultáneo con matriz de

varianza-covarianza)
4. Simultaneous RQ factor analysis with correlation matrix.(FA RQ simultáneo con matriz de correlación)

Choice ? (Elección?)

Existen cuatro tipos diferentes de PCA que se pueden hacer con esta rutina. El primero es el estándar, utilizando la matriz de varianza-covarianza derivada de su conjunto de datos. La segunda utiliza la matriz de correlación en vez de la matriz de varianza-covarianza, para encontrar los componentes principales. La tercera y cuarta opciones son variantes de los dos primeros y realizan un análisis de factores modo RQ simultáneo sobre los datos. La principal diferencia es que los análisis de factores modo RQ dan los "factor scores" no sólo para los casos sino también para las variables, lo que permite hacer lo que se conoce como un biplot. (Ver la bibliografía pertinente para obtener más información.)

Enter the name of the dataset from which principal components are to be calculated (Escriba el nombre del archivo de datos del que se calcularan los componentes principales):

Escriba el nombre del archivo de datos que desea utilizar como fuente de los componentes principales. Esto puede ser todo su conjunto de datos, o puede tratarse de un solo grupo que desee proyectar contra otros.

Luego se le ofrecerá la opción habitual de seleccionar un subconjunto de variables disponibles en su base de datos. Pulse Enter para aceptar el valor por defecto (N) si desea utilizar todas las variables del conjunto de datos. Tras la selección de las variables, si está haciendo un análisis de factores modo RQ simultáneo, se le pedirá un nombre para el conjunto de datos de scores de las variables. Estos son los scores que se grafican en un biplot como flechas y que representan la influencia de cada variable sobre la variación total del conjunto de datos.

A continuación, se le preguntará si desea obtener factor scores para sus datos. Si desea graficar sus datos en el espacio de componentes principales, deberá guardarlos, pero si lo único que le interesa es el informe del PCA (lista de eigenvalores y eigenvectores), entonces no es necesario guardar los resultados.

Luego se le ofrecerá la opción de obtener scores para otros conjuntos de datos. Esto le permite proyectar otros conjuntos de datos contra los componentes principales de su conjunto de datos original, con lo que puede comparar varios grupos en el mismo gráfico de componentes principales. Si decide hacerlo, se le preguntará por el nombre del archivo de datos que desea proyectar contra los componentes principales. A continuación, tendrá que introducir el nombre que desee para este conjunto de datos proyectado y se le preguntara si desea hacer otra proyección.

Por último, es necesario introducir un nombre para el informe. Asegúrese de incluir la extensión *.txt* de modo de poder abrirlo en un editor de texto para su visualización.

9. Canonical discriminant analysis (Análisis discriminante canónico)

Esta rutina realiza un análisis discriminante canónico (CDA) en una serie de grupos. También puede proyectar grupos adicionales en el espacio discriminante calculado para los primeros grupos.

Canonical Discriminant análisis (Análisis discriminante canónico)

Note that each of your groups must be in a separate dataset. (Tenga en cuenta que cada uno de los grupos debe ser un conjunto de datos separados).

Also note that each dataset must contain the same variables, in the same order (También tenga en cuenta que cada conjunto de datos debe contener las mismas variables, en el mismo orden).

Finally, remember that you should already have taken care of missing values in your data (Por último, recuerde que ya debería haberse ocupado de los valores faltantes)

How many groups (default=Quit)? (Cuántos grupos (default= Salir)?

Cada uno de los grupos que han de incluirse en el análisis tiene que ser un conjunto de datos separados y todos los valores faltantes deben ser calculados antes de ejecutar la rutina. Primero introduzca el número de grupos que desea incluir en el análisis. (Si en este momento usted no quiere correr el análisis, solo pulse Intro para volver al menú principal). Se le preguntará entonces por los nombres de las bases de datos que contienen los grupos (no incluya la extensión .dat). Luego se le ofrecerá la opción de seleccionar un subconjunto de las variables disponibles en los conjuntos de datos (las bases de datos deben tener todas las mismas variables en el mismo orden).

El programa luego le preguntará por un nombre para los valores de los discriminant scores que produce como salida. No incluya la extensión al ingresar este nombre. Después tendrá la opción de obtener scores para otros conjuntos de datos (asegúrese que tienen las mismas variables que el grupo original). Los scores se generan a partir de las funciones discriminantes que el programa calcula. Se le pedirá el nombre del archivo de datos para el que se desea para obtener los scores, y luego el nombre para los scores. Tiene la opción de hacerlo nuevamente.

Por último, se le pide el nombre del archivo de resultados. Este es un archivo de texto que contiene los factor loadings y los resultados del test de significación Wilk's lambda. Asegúrese de incluir la extensión .txt para poder abrirlo con un editor o procesador de texto.

10. Classify specimens using Mahalanobis distance (Clasificar casos utilizando la distancia de Mahalanobis)

Esta rutina produce una tabla de probabilidades de pertenencia a grupos basada en la distancia de Mahalanobis. Toma como entrada los grupos que se hayan definido por cualquier otro medio (gráficos bivariados, cluster analysis, datos de procedencia, etc) y genera probabilidades para la pertenencia de todos los casos a cada grupo. Esto permite recortar o redefinir los grupos sobre la base de lo que podría considerarse el núcleo de miembros de cada grupo - aquellos casos que se ajustan a su propio grupo mucho mejor que a los demás.

Al igual que con todas las rutinas que tienen más de un conjunto de datos como entrada, todos los grupos deben tener las mismas variables en el mismo orden. El cálculo de la distancia de Mahalanobis también requiere que los grupos tengan, al menos, dos muestras más que variables (cuantas más mejor). Técnicamente, sólo debería necesitar una muestra más que variables y, a veces, podrá ser que funcione, si eso es lo que usted tiene, pero por lo general se obtiene un error (*Matriz not positive definite),* lo que significa que usted tiene una matriz singular que no puede ser invertida). Si no tiene muchas muestras, será necesario realizar un análisis de componentes principales de sus datos para concentrar la mayor parte de la variación de los mismos en la menor cantidad de variables posible. A continuación, puede ejecutar esta rutina usando los conjuntos de valores de component scores y seleccionar los primeros componentes principales como sus variables.

La rutina primero le preguntará el número de grupos que desea comparar. A continuación, se le pide que introduzca el nombre de cada conjunto de datos (no incluya la extensión). Luego se le ofrecerá la opción de seleccionar un subconjunto de las variables disponibles. Si cada uno de los grupos contiene por lo menos dos muestras más que variables, no es necesario seleccionar un subconjunto a menos que se desee hacerlo. Recuerde que cuanto mayor sea la proporción muestras/variables, más confiable será el resultado.

Compute jackknife probabilities for comparing specimens
within a group to the centroid of that group (default=Y)?

Su próxima opción es calcular las probabilidades en base a los grupos tal cual (N), o calcular probabilidades "jackknife" para casos pertenecientes al grupo actualmente en

cuestión (Y). Esto elimina la muestra de la que se están calculando las probabilidades, del grupo al que pertenece, al momento de computar su probabilidad de pertenencia a ese grupo. No se afectarán las probabilidades de los casos cercanos al centro del grupo, pero es probable que para aquellos casos cerca de la orilla de un grupo, las probabilidades de pertenencia a su grupo inicial se reduzcan considerablemente. A menor relación casos/variables, esta diferencia será aún más pronunciada. Véase la bibliografía pertinente para obtener más información.

Luego se le preguntará si desea asumir homogeneidad de las matrices de varianza - covarianza para los grupos analizados. Si cree que los grupos están igualmente distribuidos en el espacio multivariado con respecto a su tamaño y forma, o si quiere tratarlos como si así fuera, entonces ingrese Y. La mayoría de las veces, o bien no sabrá o no querrá tratarlos como grupos igualmente distribuidos, en cuyo caso, deberá seleccionar la opción predeterminada (N).

Por ultimo, introduzca un nombre para el archivo de salida. Este es un archivo de texto, así que asegúrese de añadir la extensión .txt al escribir el nombre.

11. Project specimens using Mahalanobis distance (Proyectar casos utilizando la distancia de Mahalanobis)

Esta rutina realiza una función similar a la opción 10, con la diferencia de que asume que se tienen uno o más grupos bien establecidos contra los que se quiere comparar uno o más grupos de incógnitas. Mientras que la opción 10 se utiliza principalmente para perfeccionar los grupos químicos definidos en los datos, esta rutina se utiliza para comparar sus grupos con otros ya establecidos. Para obtener más información, consulte el archivo de ayuda para la opción 10 (Classify specimens using Mahalanobis distance -Clasificación de casos utilizando la distancia de Mahalanobis).

Esta rutina es muy fácil de usar. Inicialmente solicita el número de grupos de referencia que tiene. Estos son los grupos bien establecidos contra los que usted quiere comparar sus propios grupos. La rutina luego le pregunta el número de archivos con casos para comparación (sus grupos). Es importante tener una relación casos/variables tan alta como sea posible en cada grupo para el que se quiera calcular probabilidades (también para los grupos de referencia). Si no tiene al menos dos casos más que variables en cualquiera de sus grupos, tendrá que reducir el número total de variables, ya sea seleccionando un subconjunto o, mejor, realizando un PCA de sus datos (y/o los datos de referencia) para tener la mayor cantidad de variación total en el menor número de variables posible. Luego puede utilizar los component scores como input para esta rutina y seleccionar los primeros componentes principales como el subconjunto de las variables a utilizar para el cálculo de probabilidades.

Una vez que introduzca los nombres de sus conjuntos de datos, se le pedirá que seleccione un subconjunto de variables (si así lo desea). Por último, es necesario ingresar un nombre para el archivo de texto de salida (no olvide incluir la extensión). Cuando mire este archivo, aparecerá cerca del final algo como lo siguiente:

Summary of Probabilities for Specimens in the file MYGP1 (Resumen de las probabilidades para los casos en el archivo MYGP1

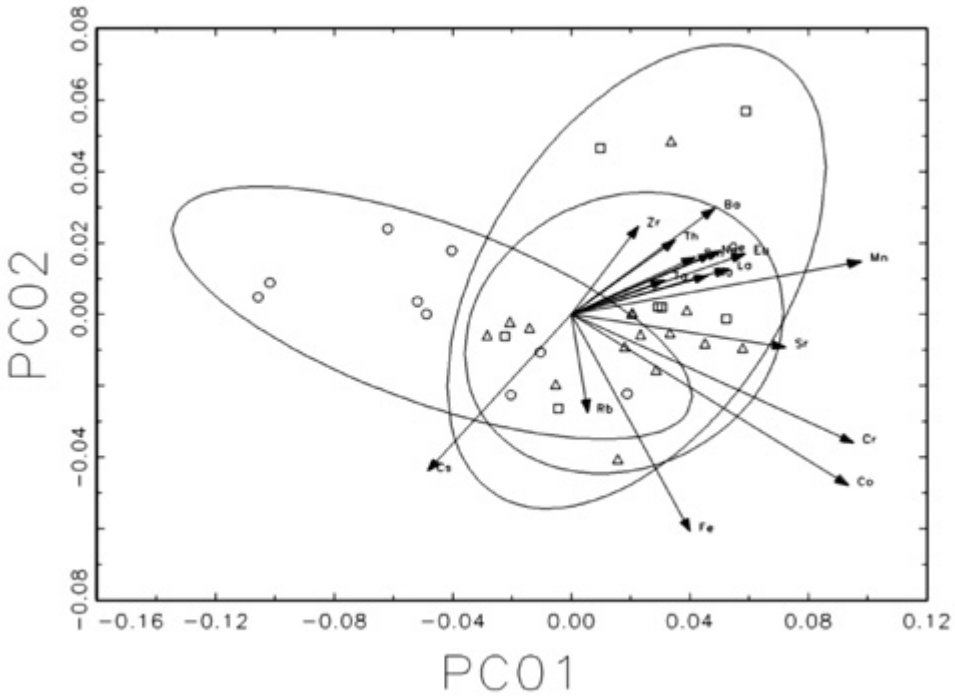
	Probability Cutoff Values:						
Group:	*0.01*	*0.10*	*1.00*	*5.00*	*10.00*	*20.00*	*100.00*
REF1	*8*	*14*	*23*	*37*	*44*	*58*	*94*
REF2	*1*	*3*	*5*	*10*	*14*	*18*	*94*

Las cuentas que se muestran representan el número de casos del archivo que aparece en el encabezado (en este caso, MYGP1), que tienen probabilidades de inclusión en los grupos de referencia, en la columna de la izquierda debajo de la lista de los valores de corte de probabilidad. Una forma de usar esta información sería seleccionar un valor de corte para la pertenencia a un grupo: Toda muestra con un valor inferior al de corte se considerará fuera de ese grupo y se colocará en una categoría sin asignar. A menudo tendrá casos con

muy poca probabilidad de pertenencia a cualquiera de sus grupos de referencia, y la única forma de tratar con ellos es ubicarlos en una categoría sin asignar. Todo se reduce a decidir cual debe ser el valor de corte. Por favor, consulte la literatura relevante para la orientación en la materia.

12. 2D scatterplots (Gráficos de dispersión en dos dimensiones)

Esta opción le permite producir gráficos bivariados de conjuntos de datos Gauss. Todos los grupos que se desea graficar deben estar en distintas bases de datos y todos los conjuntos de datos deben tener las mismas variables en el mismo orden. La rutina permite graficar puntos, elipses de confianza, y producir lo que se conoce en la literatura como un "biplot".



Un biplot puede producirse a partir de la salida de un PCA modo RQ simultáneo. Las flechas que se ven en el gráfico corresponden a los scores de las variables. Los gráficos bivariados normales también pueden producirse con o sin elipses de confianza. También puede graficar una elipse de confianza para un grupo si quiere ver si hay superposición de un conjunto desconocido de casos con algún grupo conocido.

La rutina comienza preguntando por el intervalo de confianza que desea para todas las elipses que se graficarán. Introduzca un número entre 0 y 1. Por lo general, se utilizan intervalos de confianza del 90 o 95%, por lo que debe entrar 0,9 o 0,95 para estas elipses, respectivamente.

A continuación, se le pedirá que introduzca el número de grupos que desea graficar. Antes de ejecutar esta rutina recuerde que cada grupo tiene que ser un conjunto de datos separado. Para el gráfico de arriba, se optó por graficar cuatro grupos (tres grupos químicos más el conjunto de datos que contiene los scores de las variables).

Se le pedirá el nombre del primer conjunto de datos. Después de que lo ingrese (no importa el orden en el que introduzca los nombres de los conjuntos de datos), podrá seleccionar las variables que desea graficar. La variable x que se elija estará en todos los gráficos y las variables y se representarán gráficamente en función de la variable x, a partir de la variable y que se especifique (si desea ver la totalidad de los gráficos contra una única variable x, elija que y comience con la primer variable).

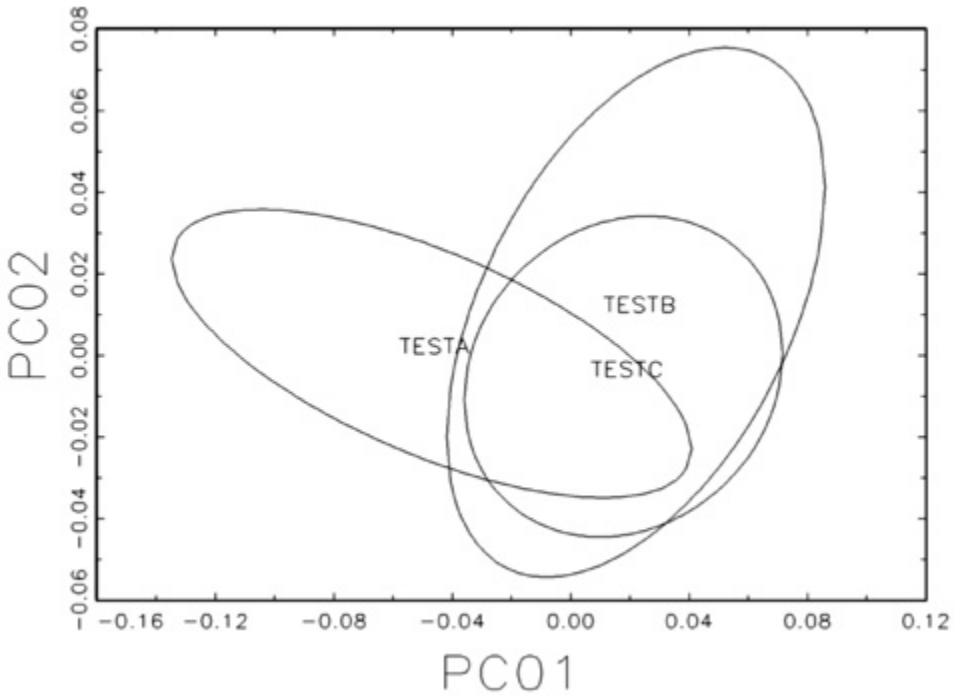
El programa luego le pide que seleccione un símbolo a utilizar al graficar el primer conjunto de datos. Aún si no piensa graficar los puntos individuales, igual tiene que seleccionar un símbolo - no importa cuál elija.

Luego, se le pregunta si desea asumir un tamaño de muestra muy grande al calcular la elipse de confianza. Si tiene más de alrededor de 120 casos representados en su conjunto de datos, no importa si usted decide hacerlo o no. Sólo habrá realmente una diferencia para tamaños de muestra pequeños (menos de 30 casos). Si tiene menos de 30 muestras en un grupo y no opta por asumir un gran tamaño de muestra, la elipse obtenida será algo más grande de lo que sería del otro modo. Esto se debe a que se toma en cuenta el tamaño de la muestra (en términos de cuan representativa puede ser la muestra de la población en su conjunto), por lo que muestras más pequeñas tienen mayores elipses de confianza para un determinado nivel de confianza estadística.

El programa entonces volverá a correr a través de la selección de los datos, tipo de símbolo, tamaño de muestra asumido para el número de grupos especificado primero. Tendrá que ver cuales cuál grupo se grafica con qué símbolo, dado que el programa no genera una leyenda para el grafico final.

Una vez que haya terminado con esto, deberá especificar lo que quiere hacer con cada uno de los grupos que ha introducido. En el ejemplo anterior, los conjuntos de datos fueron graficados utilizando la opción # 2 (gráfico y elipse) y los variable scores se graficaron con la opción # 7 (label points and vector from origin). Al seleccionar esta última opción, también se tiene la opción de graficar las líneas con flechas, como se muestra en el gráfico de arriba.

Su opción final para la preparación del grafico es si quiere o no los nombres de los grupos en el centro de cada uno. A menos que esté graficando sólo elipses, esto puede ser problemático, ya que puede ser difícil de leer si hay representados muchos puntos. El siguiente gráfico utiliza los mismos datos que el que se muestra más arriba, pero ilustra la forma en que se ven los nombres de los grupos, cuando sólo se grafican las elipses.



Los nombres de los archivos de datos se muestran en los centros de sus correspondientes elipses. Se puede ver que si también se hubiesen graficado puntos y vectores de variables, habría sido muy difícil de leer.

Luego el programa da la opción de mostrar todos los gráficos a la vez o de a uno. Windows XP tiene una característica que permite que varias instancias del mismo programa se agrupen en la barra de herramientas, por lo que resulta más fácil gestionar los gráficos si se decide hacer todos a la vez. Las versiones anteriores de Windows no disponen de esta función y dependiendo de la cantidad de memoria que tiene instalada en su máquina,

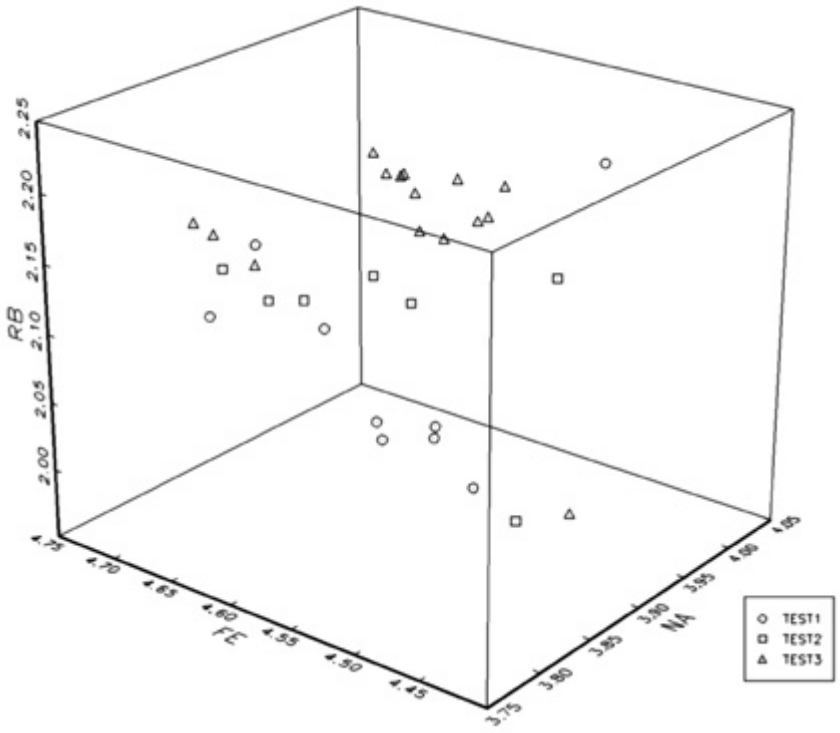
puede encontrarse con problemas si decide hacer todos los gráficos a la vez. La opción de circular a través de los gráficos puede ser la mejor elección en este caso.

Por último, ¿qué se puede hacer con los gráficos cuando se los tiene dibujados? En primer lugar, deberá familiarizarse con el TKF File Viewer leyendo el Readme en la opción Help de la barra del menú de aplicaciones. Luego, puede hacer copias (bitmap copy) de los gráficos (después de la conversión de colores a negro sobre un fondo blanco - ver View/ | Options en el TKF File Viewer) y pegarlos en Photoshop. Si desea realmente manipular los gráficos, puede elegir copiar (metafile copy) y pegar el gráfico en Illustrator (los productos de Adobe mencionados son sólo a título de ejemplo - cualquier aplicación de gráficos funciona igualmente bien. Por favor, no nos pregunten cómo utilizarlos). Una vez importados en el programa gráfico, la sorpresa desagradable es que el gráfico está, en realidad, compuesto por una multitud de líneas –el texto, por ejemplo, no es texto sino que está armado con segmentos de línea individuales. Esto no significa que la manipulación de los gráficos sea imposible, pero lleva un tiempo acostumbrarse. La clave está en la experiencia que se tenga con el programa gráfico disponible.

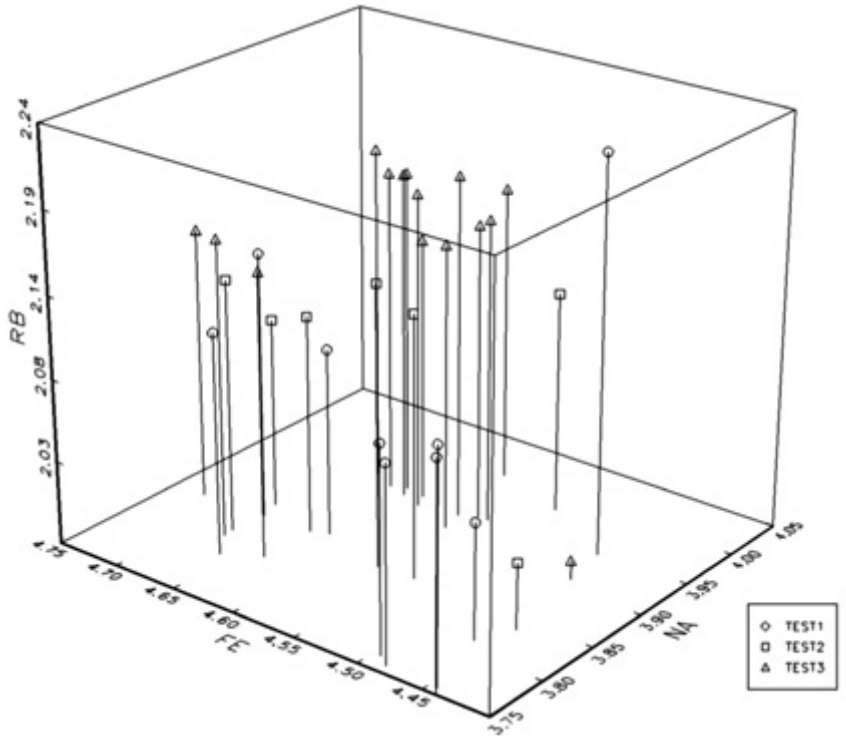
Si desea guardar un gráfico en formato TKF para su uso posterior, especifique un nombre que no tenga la forma graphic*.tkf, porque esos son los nombres de los archivos de gráficos temporales creados durante el funcionamiento normal de esta rutina. Los archivos con este tipo de nombre se borrarán al volver al menú principal.

13. 3D scatterplots (Gráficos de dispersión 3D)

Esta rutina le permite crear gráficos de dispersión en tres dimensiones (gráficos trivariados). Los graficos en 3D no son de uso común en las publicaciones, ya que tienden a ser difíciles de leer, por lo que no ofrecemos muchas opciones de modificación de los mismos. Esta rutina está destinada principalmente para su uso como herramienta de exploración de datos. El siguiente es un ejemplo de lo producido con esta rutina:



Se trata de un simple grafico trivariado de tres grupos de datos (estos gráficos incluyen una leyenda). La rutina funciona de manera semejante a la rutina de gráficos de dispersión 2D, aunque las opciones para cosas como elipses, etiquetas de muestras y grafico de variable scores no están disponibles. Sin embargo existe una opción que permite incluir proyecciones al plano XY:



La única otra diferencia entre esta rutina y la de gráficos de dispersión 2D es en la selección de variables para graficar. Debe seleccionar tres variables a graficar y las variables x e y se mantienen constantes para todos los gráficos que se dibujan. Para más información, por favor consulte el archivo de ayuda para la rutina de gráficos de dispersión 2D.

14. Convert all v89 files to v96 format in current directory (Convertir todos los archivos de formato v89 a v96 en el directorio actual)

Gauss 5.0 utiliza un formato de archivo de datos diferente que las versiones anteriores de Gauss (como Gauss 3.4 o 3.6). Si tiene alguno de estos archivos de formato viejo (v89), que consiste de un archivo .dat y un archivo .dht, puede usar esta opción para convertirlos al nuevo formato (v96), que emplea sólo un nuevo tipo de archivo .dat. Tenga en cuenta que esta es una conversión sin retorno (que no se puede deshacer) y que sus antiguos archivos v89 se borrarán.

15. Help files (archivos de ayuda)

Al seleccionar esta opción se abren los archivos de ayuda que usted está leyendo ahora. El sistema de Ayuda puede permanecer abierto mientras trabaja en Gauss, para que pueda remitirse a las instrucciones según sea necesario.

El sistema de ayuda es el mismo que la mayoría de las demás ayudas en línea que ofrece Windows. La ventana mantiene, de una sesión a la siguiente, cualquier cambio (en términos de tamaño) que le haya hecho. Pulsando el botón derecho dentro de la ventana se tiene la opción de imprimir el contenido, si necesita una copia impresa.

16. Exit program (Salir del programa)

Esta opción le permite detener la ejecución de las rutinas estadísticas MURR del menú. Esto es útil cuando se quiere dejar de usar Gauss, ya que el programa no se detendrá por sí solo cuando se encuentra a la espera de una entrada. Una solución es cerrar Gauss haciendo clic en la "x" en la esquina superior derecha del marco de la ventana GUI, luego seleccionar OK cuando se le pregunta si realmente desea salir y luego pulsar Enter para

realmente detener el programa.

Como se discutió en la sección Primeros pasos de la introducción de este documento de ayuda, si desea reiniciar el menú una vez que haya elegido salir, puede escribir run sarmenu.gcg (si se encuentra en el directorio raíz de la aplicación Gauss) y pulsar la tecla Enter, o si ya tiene sarmenu.gcg en el Program List, entonces hacer clic en Run (Ejecutar).

Información de Contacto

Si tiene otras preguntas que no pueden ser respondidas por estos archivos de ayuda, no dude en ponerse en contacto con Mike Glascock en MURR.

Dr. Michael D. Glascock
Research Reactor Center
University of Missouri
Columbia, MO 65211
Email: GlascockM@missouri.edu
Phone: (573) 882-5270
FAX: (573) 882-6360
Laboratory Home Page: http://archaeometry.missouri.edu