AOML NOAA

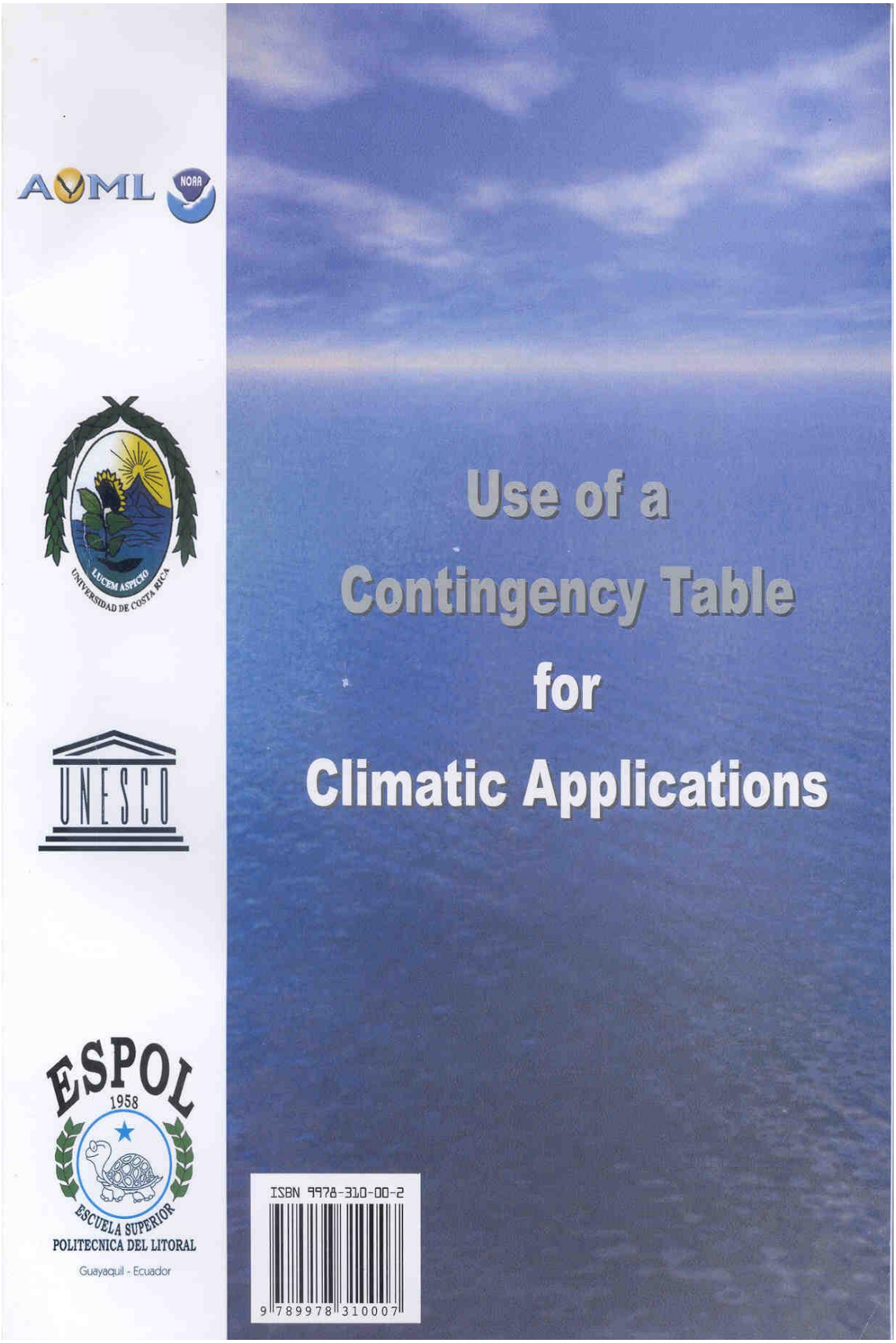
LUCEM ASPICIO
UNIVERSIDAD DE COSTA RICA

UNESCO

ESPOL
1958
ESCUELA SUPERIOR
POLITECNICA DEL LITORAL
Guayaquil - Ecuador

Uso de una Tabla de Contingencia para Aplicaciones Climáticas

ISBN 9978-310-00-2
9 789978 310007


AOML
NOAA
LUCEM ASPICIO
UNIVERSIDAD DE COSTA RICA
UNESCO
ESPOL
1958
ESCUELA SUPERIOR
POLITECNICA DEL LITORAL
Guayaquil - Ecuador
Use of a
Contingency Table
for
Climatic Applications
ISBN 9978-310-00-2
9 789978 310007

Uso de una Tabla de Contingencia para Aplicaciones Climáticas

Eric J. Alfaro y F. Javier Soley

Escuela de Física y Centro de Investigaciones Geofísicas
Universidad de Costa Rica

David B. Enfield

Physical Oceanography Division
Atlantic Oceanographic and Meteorological Laboratory
National Oceanic and Atmospheric Administration

Índice

Manual del Usuario

1. Introducción

Desde 1997 se han venido realizando en distintas partes de Latinoamérica los llamados Foros Regionales de Predicción Climática (conocidos como RCOFs por sus siglas en inglés), financiados por diversas agencias internacionales y con la asistencia de distintas entidades como el Comité Regional de Recursos Hidráulicos o CRRH en Centroamérica.

Estos foros, generalmente reúnen a los representantes de los servicios meteorológicos e hidrológicos, así como a los miembros de la comunidad científica y académica, que trabajan en la elaboración de las perspectivas climáticas regionales y locales. El objetivo de estos foros es el de usar la experiencia climática nacional para elaborar una perspectiva climática de consenso regional, generalmente de precipitación, de los próximos meses y que además se presente en una forma útil para las distintas agencias involucradas. La metodología recomendada para los mismos es simple, en la cual las probabilidades de los distintos terciles de la precipitación o temperatura, como predictante, se extraen de una tabla de contingencia de dos variables en donde se usa algún índice climático como predictor, p.e. el Índice de Oscilación del Sur (IOS) o índices asociados al promedio de la Temperatura Superficial del Mar (TSM) en alguna región como por ejemplo Niño 3 o el Atlántico Tropical Norte (ATN). Esta perspectiva se integra luego regionalmente para ayudar a los distintos servicios meteorológicos en sus diversas actividades, así como también a los tomadores de decisión y grupos de interés involucrados.

Con el fin de mejorar estos foros, la comunidad científica y académica ha discutido algunos problemas que han enfrentado estos foros regionales y las posibles formas en las cuales los resultados de sus investigaciones pueden ayudar con estos procesos. A raíz de esto, se ejecutó un proyecto financiado por la Oficina de programas Globales de la Agencia para la Atmósfera y el Océano de los Estados Unidos de América (OGP-NOAA), apoyado por el Centro de Investigaciones Geofísicas de la Universidad de Costa Rica y el Laboratorio de Meteorología y Oceanografía para el Atlántico de la NOAA (de aquí en adelante citado como Proyecto NOAA-UCR). Este proyecto de extensión pretende atender algunas de las necesidades que en materia de predicción climática tienen los servicios meteorológicos nacionales, con el fin de que su participación en los RCOFs sea más eficiente y coordinada entre los países participantes. Dentro de los problemas detectados se puede mencionar que al no tener una metodología estandarizada, las diferentes contribuciones nacionales a la predicción regional no son uniformes y algunas veces no son consistentes físicamente, teniéndose resultados en países vecinos diametralmente opuestos a lo largo de sus fronteras. Por otra parte, el fundamento estadístico sobre la metodología de los terciles parece no serles muy familiar a algunos de los participantes, por lo que sus perspectivas climáticas se basan en ocasiones en evaluaciones subjetivas. Algunas de las razones de lo anterior son: i) los recursos de algunas instituciones están limitados al quehacer diario y pueden asignar muy poco de su presupuesto a actividades de investigación y capacitación; además, ii) hasta ahora ha

habido muy pocas oportunidades de entrenamiento sobre estos conceptos para los participantes en estos foros.

Tomando en cuenta lo anterior, el proyecto NOAA-UCR antes descrito ha creado una serie de programas amigables para facilitar la implementación del pronóstico climático usando el concepto de los terciles. Los programas usan interfaces gráficas con el usuario (GUIs, por sus siglas en inglés) y versiones preliminares del software han sido distribuidas al personal de los servicios meteorológicos por medio de un CD-ROM, el cual acompaña a este manual y que puede ser usado en sesiones de entrenamiento durante los RCOFs. Se incluye una biblioteca con una serie de funciones especializadas, compiladas bajo ambiente MATLAB. El objetivo principal es el análisis cuantitativo y categórico de dos variables. Esto se logra por medio de la construcción de una tabla de contingencia para evaluar el grado de asociación entre las parejas de los dos conjuntos de datos así como de su relación condicional de probabilidad. Las funciones de Matlab se compilaron luego para el ambiente Windows independientes de la plataforma Matlab[1]. Esta manual es para la última versión del software. La historia de las distintas versiones del software se describe en el Apéndice.

El programa principal `wtc.exe` o `wtceng.exe`, (versiones del software en español e inglés respectivamente) puede ser usado en dos formas. La primera para hacer el análisis exploratorio entre dos variables que se suponen correlacionadas, positiva o negativamente, en el tiempo (p.e. La precipitación en San José vs. el IOS) y la segunda para realizar la validación de un pronóstico, o sea, ver como se contrastan los valores observados con respecto a los pronosticados. El otro programa es `wfcc.exe` o `wfcceng.exe` (nuevamente, versiones del software en español e inglés respectivamente), el cual realiza la correlación cruzada sesgada entre dos variables y ayuda a escoger un predictor apropiado.

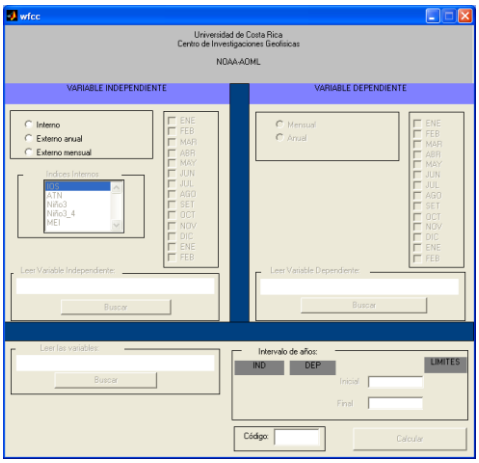
2. Ejecución del programa

Los programas son ejecutables bajo ambiente Windows. Como primer paso el usuario debe copiar el directorio `exever` del CD a su disco duro. En este directorio se encuentran los programas ejecutables `wfcc.exe` y `wtc.exe` (o `wfcceng.exe` y `wtceng.exe`). Seguidamente el usuario debe crear un atajo (`shortcut`) de estos programas en el escritorio (`desktop`) de su computadora, esto evitará que se borre accidentalmente alguna de las bibliotecas incluidas en `exever` necesarias para el funcionamiento de los programas.

El primer programa que discutiremos es `wfcc.exe` o `wfcceng.exe`, el cual realiza la correlación cruzada sesgada entre dos variables para los rezagos –1, 0 y +1. Este programa es de mucha ayuda para la identificación de predictores. Las dos variables sobre las que se realiza el análisis deben ser secuencias temporales anuales, aunque las secuencias de entrada al programa pueden ser mensuales o anuales. Se interpreta que para

[1] Existe también una versión preliminar que trabaja en ambiente MS-DOS, esta se incluye también en el CD, así como el respectivo Manual del Usuario en español, refiérase al Apéndice para mas detalles.

rezagos negativos la variable independiente antecede a la variable dependiente. Al ejecutar este programa se despliega la siguiente interfáz gráfica al usuario:

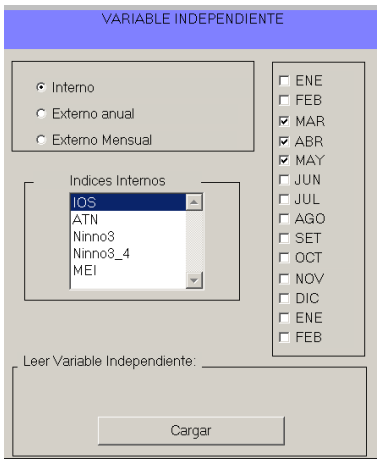


Las variables que serán analizadas deben estar contenidas en sendos archivos ASCII o de texto simple, en los cuales los valores de las filas de cada columna deben estar separados por uno o más espacios en blanco. La primera columna es un índice secuencial y la segunda columna la variable independiente o dependiente[2]. Las líneas que se consideren comentarios deben iniciarse con un signo de % y los datos faltantes deben codificarse apropiadamente con un valor numérico, p.e. -9999.

Esta interfaz gráfica le permite al usuario importar los datos de distintas maneras. La variable independiente (predictor) puede ser leída de tres formas. Si el usuario selecciona `Interno`, esto quiere decir que va a usar uno de los índices climáticos que tiene incorporados el programa, para lo cual no se requiere que el usuario prepare un archivo de antemano. Se debe seleccionar alguno de ellos. Los cinco archivos de texto (`*.txt`) correspondientes a estos índices, IOS, ATN, Niño3, Niño 3.4 y MEI, se encuentran en el directorio `c:\exever\indices`. Posteriormente el usuario debe seleccionar la estación climática del año que desea analizar de este índice, p.e. seleccionar `IOS` y luego `MAM (marzo, abril, mayo)`[3], marcando los respectivos meses en la columna de la derecha,

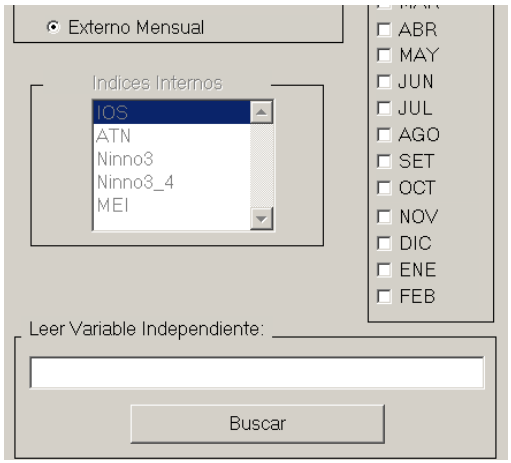
[2] Se incluye también una macro en excel que transforma los datos mensuales del formato usual (ie Una matriz de N años y 13 columnas: año, Ene, Feb, ..., Dic) a un vector de datos mensuales en el formato usado por `wfcc.exe` o `wfcceng.exe`. El archivo se llama mattira.xls

[3] En este mismo directorio se incluyen archivos de texto para los índices Niño1+2, PDO y AMO. Estos pueden ser usados por el programa por medio de la opción Externo Mensual, la cual se describe posteriormente en esta sección.

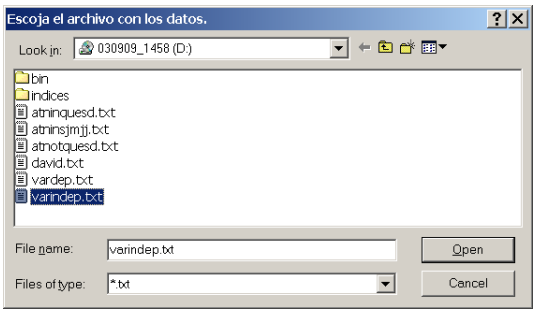


Luego el usuario seleccionará `Cargar`, para pasar estos datos anuales del predictor a la memoria.

Si el usuario selecciona `Externo Mensual`, quiere decir que se usará un índice climático no contemplado en el menú y que se encuentra en un archivo texto de dos columnas como se mencionó anteriormente, la primera corresponde al eje temporal y la segunda al índice climático. La búsqueda del archivo a ser usado debe hacerse a través del botón `Buscar`,



después de lo anterior, sigue seleccionar el archivo a usar haciendo `clic` en `abrir` (`open`),



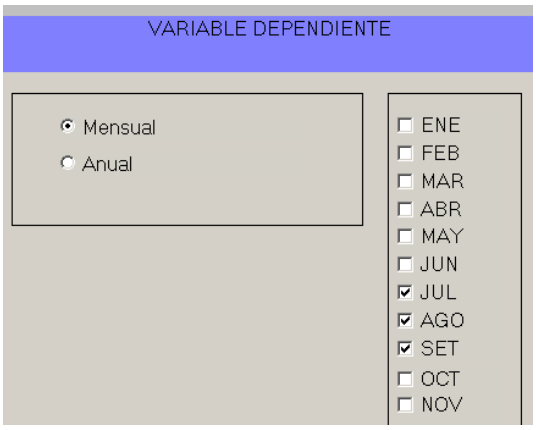
Al igual que en el caso anterior, debe seleccionar los meses correspondientes a la estación climática que desee analizar para la variable independiente y el programa crea la secuencia temporal anual a partir de los datos mensuales. Note que si el usuario coloca el directorio exever con una dirección distinta a c:\exever, los índices desplegados usando la opción Interno no serán hallados por el programa, pero aún así pueden ser usados por medio de la opción Externo mensual.

Cuando el usuario oprime `Externo Anual` en el menú de la `Variable Independiente`, las variables anuales que serán analizadas deben estar contenidas en un archivo ASCII o texto simple de tres columnas. La primer columna es un índice secuencial (en la mayoría de los casos corresponde al año, pero no necesariamente), la segunda columna se considera la variable independiente y la tercer columna es la variable dependiente. Cabe destacar que en todos los casos las líneas que se consideren comentarios deben iniciarse con un signo de % y como se mencionó anteriormente, los datos faltantes deben codificarse apropiadamente con un valor numérico. Por ejemplo[4]:

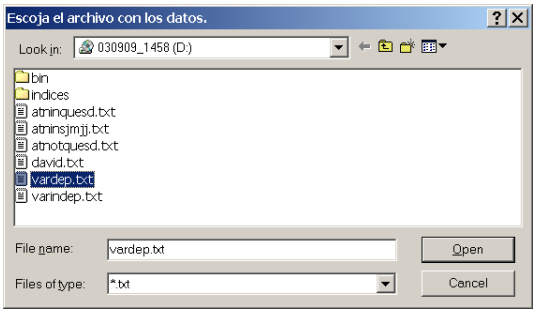
```
% Año             IOS, JJA(-1)      Quepos, IELL
  1.9410000e+003 -1.7733333e+001  1.9000000e+001
  1.9420000e+003 -1.8033333e+001  2.1000000e+001
  1.9430000e+003  3.7666667e+000  2.7000000e+001
  1.9440000e+003  9.3333333e-001  2.2000000e+001
  1.9450000e+003 -3.1666667e+000  2.3000000e+001
  1.9460000e+003  7.8333333e+000  2.4000000e+001
  1.9470000e+003 -8.0666667e+000           -9999
  1.9480000e+003  6.4000000e+000           -9999
  1.9490000e+003 -2.7333333e+000  2.4000000e+001
```

La lectura de la variable dependiente puede hacerse de dos formas. Como primer paso se debe seleccionar `Mensual` o `Anual`, dependiendo de la naturaleza de los datos. Si se selecciona `Mensual` nuevamente se activará la selección de los meses que permiten escoger una estación climática del año específica a ser usada en el análisis, p.e. JAS,

[4] El formato de entrada de los datos no tiene que ser necesariamente exponencial como en este ejemplo, vea la sección 4 para más detalles.

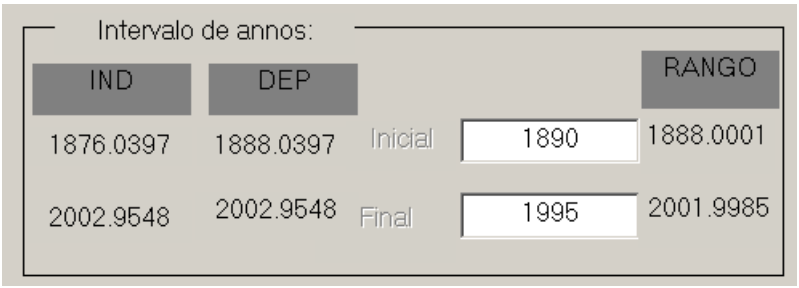


al igual que en el caso anterior, se debe usar el botón `Buscar` para seleccionar el archivo del análisis,



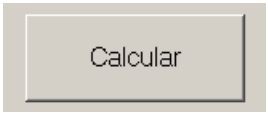
La opción `Anual` de la `Variable Dependiente` permite seleccionar una serie de datos anuales, en donde el archivo texto a importar contiene dos columnas: la primera con el índice secuencial (años usualmente) y la segunda con los datos del parámetro a analizar.

Una vez que se tengan las dos variables, independiente y dependiente, en la memoria, el usuario debe seleccionar el intervalo del índice secuencial con el cual desea trabajar. Para ello se despliegan en el extremo izquierdo de la interfaz gráfica los valores máximos y mínimos de las variables usadas, así como el intervalo común entre ellas. El periodo que el usuario seleccione para trabajar debe ser un subconjunto de este intervalo común.

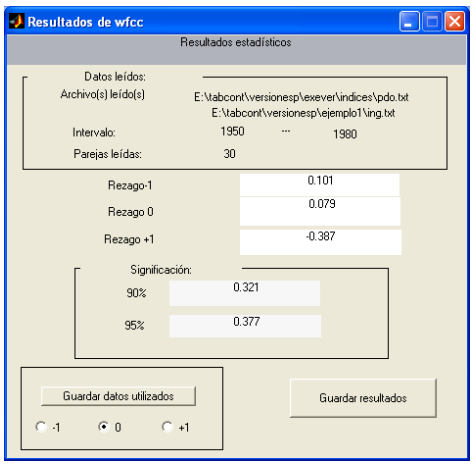


Este análisis permite la existencia de un número razonable de datos faltantes. Si los hay el usuario debe escribir en `Código` el valor correspondiente al código del dato faltante que está usando en su análisis, o escribir nh si no hay datos faltantes.

A continuación y para efectuar el análisis, el usuario debe oprimir `Calcular`,



lo que desplegará la pantalla con el análisis de la función de correlación cruzada. La interpretación y uso de estos resultados será explicado luego en esta sección.

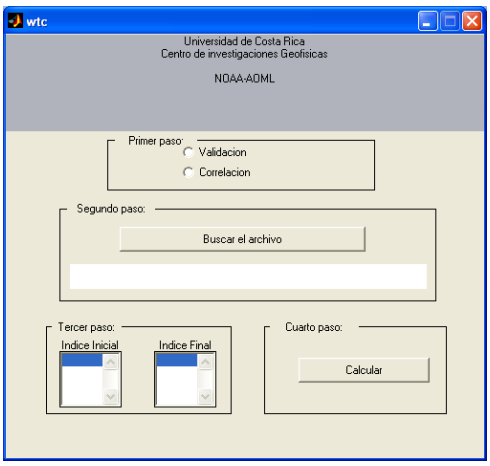


Por último si desea guardar los resultados de su análisis en un archivo texto se oprime `Guardar resultados`. Si el usuario desea guardar los datos utilizados en un archivo de texto para el posterior análisis de contingencia de debe presionar Guardar datos utilizados, marcando además el rezago apropiado para el predictor, es decir –1 para el rezago –1, el año antes, 0 para el rezago 0, el mismo año o +1 para el rezago +1 o el año después.

El segundo de los programas es `wtc.exe` o `wtceng.exe` el cual construye la tabla de contingencia entre dos variables. Al igual que en el caso anterior, las variables que serán analizadas deben estar contenidas en un archivo ASCII en donde la primer columna es un índice secuencial, la segunda columna se considera la variable independiente y la tercer columna es la variable dependiente. Las líneas que se consideren comentarios deben iniciarse con un signo de %. Si en una fila ocurre uno o dos datos faltantes, estos se eliminan del análisis anteponiendo el mismo símbolo. Por ejemplo:

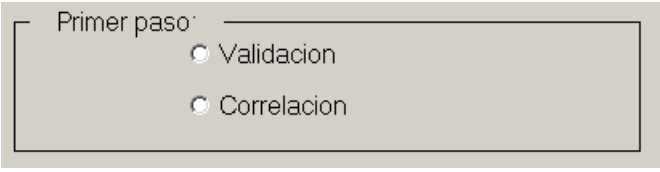
```
% Año             IOS, JJA(-1)      Quepos, IELL
  1.9410000e+003 -1.7733333e+001  1.9000000e+001
  1.9420000e+003 -1.8033333e+001  2.1000000e+001
  1.9430000e+003  3.7666667e+000  2.7000000e+001
  1.9440000e+003  9.3333333e-001  2.2000000e+001
  1.9450000e+003 -3.1666667e+000  2.3000000e+001
  1.9460000e+003  7.8333333e+000  2.4000000e+001
% 1.9470000e+003 -8.0666667e+000   NaN (o -9999)
% 1.9480000e+003  6.4000000e+000   NaN (o -9999)
  1.9490000e+003 -2.7333333e+000  2.4000000e+001
```

Al ejecutar este programa de despliega la siguiente interfaz gráfica al usuario,

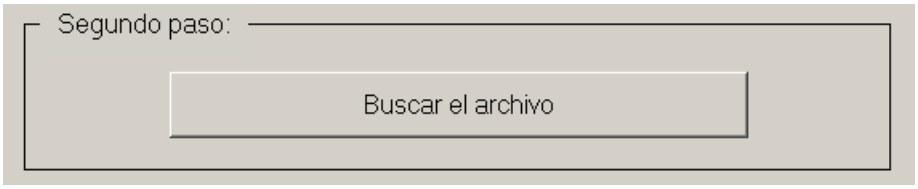


Una vez desplegada esta interfaz se siguen los siguientes pasos:

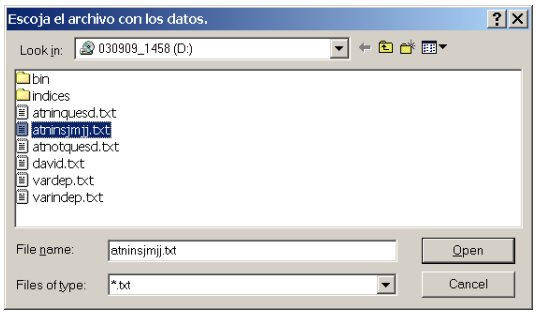
1) Seleccionar Validación, para realizar una validación cruzada y contrastar el comportamiento de datos que son producto de un pronóstico con respecto a los datos observados, o seleccionar Correlación para realizar el análisis exploratorio entre dos variables que se supone correlacionadas *positivamente* o *negativamente* en el tiempo (p.e. La precipitación en San José en MJJ vs. el IOS en DEF).



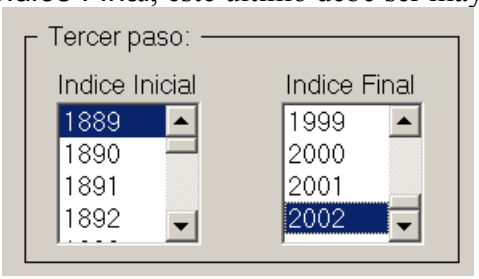
2) Como segundo paso se debe seleccionar el archivo del análisis oprimiendo `Buscar el archivo`,



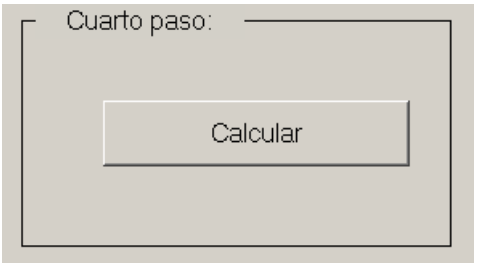
Una vez seleccionado, se presiona `abrir` u `open` para pasarlo a memoria.



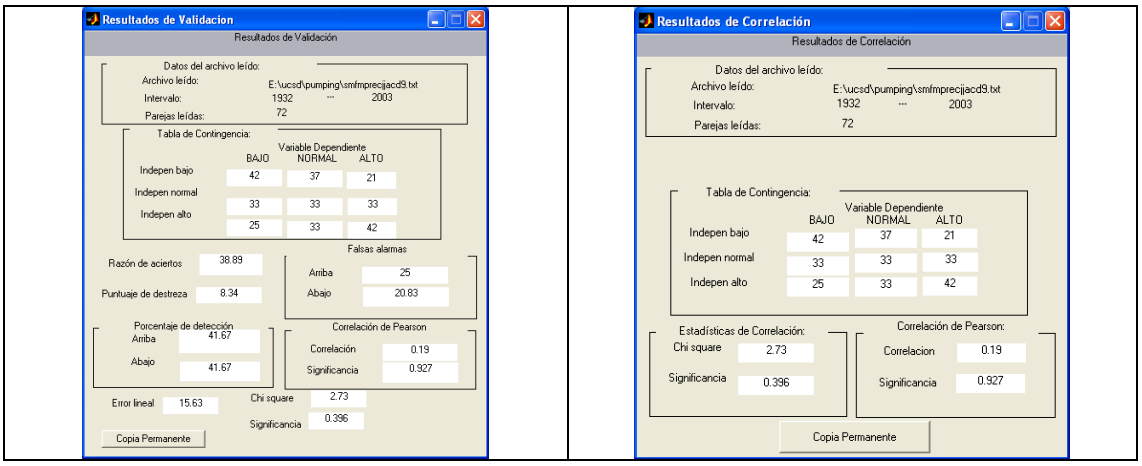
3) Como tercer paso se debe escoger el período de análisis haciendo clic en Índice Inicial e Índice Final, este último debe ser mayor que el primero.



4) En este momento se puede realizar el análisis contingente presionando `Calcular`.

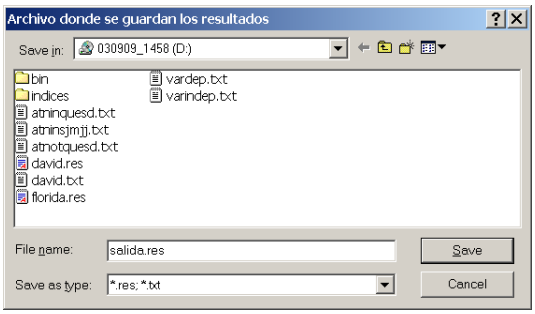


En la pantalla se desplegará la tabla de contingencia y los diferentes estadísticos del análisis, los cuales serán explicados posteriormente en la sección 3. Su formato dependerá de la opción que el usuario tomara en el primer paso, es decir, `Validación` o `Correlación`.

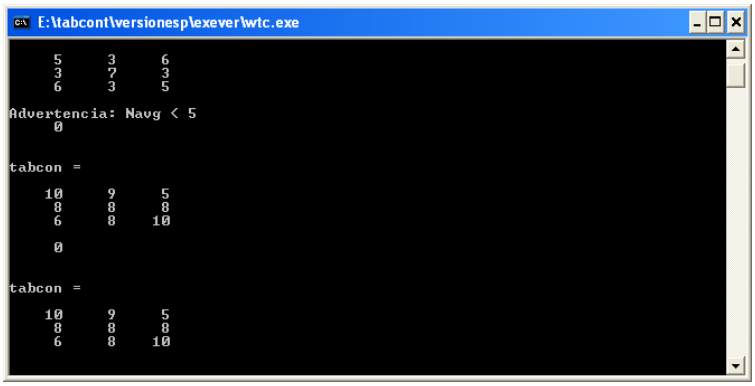


5) El quinto y último paso es opcional. En este paso se guardan los resultados del análisis en un archivo de texto o ASCII. Para hacerlo haga clic en Copia Permanente.

y digite el nombre del archivo donde quiere guardar sus resultados.



Note que puede elegir la carpeta donde guardará su archivo de resultados. Los resultados de la tabla de Contingencia en frecuencias absolutas se despliegan en la pantalla del MS-DOS y se advierte al usuario si el número de pares leídos es menor de 45 o sea, un número promedio de pares por celda menor a cinco.



3. Aspectos teóricos del programa e interpretación de los resultados.

3.1 Función de correlación cruzada.

El paso clave para la aplicación de la tabla de contingencia es la escogencia de las variables independientes y dependientes que tengan una relación predictiva significativa. En otras palabras, queremos que ellas estén correlacionadas. Una herramienta matemática que nos permite cuantificar el comportamiento común (o el grado de "información" común) entre esas dos variables es la función de correlación cruzada, que se define como:

$$c_{xy}[k] = \frac{1}{N}\sum_{t=1}^{N-k}\frac{\left(x[t]-\bar{x}\right)\left(y[t+k]-\bar{y}\right)}{\sigma_x\sigma_y} \qquad k = 0,1,2\ldots,$$

$$c_{yx}[k] = \frac{1}{N}\sum_{t=1}^{N-k}\frac{\left(y[t]-\bar{y}\right)\left(x[t+k]-\bar{x}\right)}{\sigma_x\sigma_y} \qquad k = 0,1,2\ldots,$$

donde $\bar{x}$ y $\bar{y}$ son las medias muestrales, y σ_x y σ_y las desviaciones estándar de las series *x* y *y*. Llamando con *x'* y *y'* a los residuos normalizados de las series podemos simplificar la ecuación a

$$c_{xy}[k] = \frac{1}{N}\sum_{t=1}^{N-k} x'[t]y'[t+k] \qquad k = 0,1,2\ldots,$$

$$c_{yx}[k] = \frac{1}{N}\sum_{t=1}^{N-k} y'[t]x'[t+k] \qquad k = 0,1,2\ldots,$$

De estas ecuaciones notamos los siguientes puntos de interés:

a) Llamemos al producto *x'[t] y'[t+k]* el producto de *x* con *y* adelantada *k* tiempos de muestreo. Entonces $c_{xy}[k]$ es el valor medio de los productos de *x'* con *y'* adelantada *k* tiempos de muestreo. De la misma forma $c_{yx}[k]$ es el valor medio de los productos de *y'* con *x'* adelantada *k* tiempos de muestreo. Por ejemplo: supongamos que *x'* son los residuos del IOS y *y'* los residuos de una secuencia de precipitación mensual. $c_{xy}[0]$ es el valor promedio de los productos IOS[Ene] PREC[Ene], IOS[Feb] PREC[Feb], IOS[Mar] PREC[Mar], etc. $c_{xy}[1]$ es el valor promedio de los productos IOS[Ene] PREC[Feb], IOS[Feb] PREC[Mar], IOS[Mar] PREC[Abr], etc. $c_{xy}[2]$ es el valor promedio de los productos IOS[Ene] PREC[Mar], IOS[Feb] PREC[Abr], IOS[Mar] PREC[May], etc.
b) Supongamos que *x'[t]* y *y'[t+k]* no tienen un comportamiento común. Es decir, cuando *x'[t]* obtiene un valor positivo, *y'[t+k]* indistintamente obtiene valores positivos o negativos y cuando *x'[t]* obtiene un valor negativo, *y'[t+k]* también indistintamente obtiene valores positivos o negativos. Entonces el valor promedio tiende a anularse y decimos que las dos series no están correlacionadas en rezago *k*.
c) Supongamos que *x'[t]* y *y'[t+k]* si tienen un comportamiento común. Es decir, cuando *x'[t]* obtiene un valor positivo, *y'[t+k]* preferentemente obtiene

valores positivos y cuando *x'[t]* obtiene un valor negativo, *y'[t+k]* preferentemente obtiene valores negativos. Entonces el valor promedio es positivo y decimos que las dos series están correlacionadas positivamente en rezago *k*.

d) Supongamos que *x'[t]* y *y'[t+k]* tienen un comportamiento común y cuando *x'[t]* obtiene un valor positivo, *y'[t+k]* preferentemente obtiene valores negativos y cuando *x'[t]* obtiene un valor negativo, *y'[t+k]* preferentemente obtiene valores positivos. Entonces el valor promedio es negativo y decimos que las dos series están correlacionadas negativamente en rezago *k* o que están anticorrelacionadas en rezago *k*.

e) La definición de $c_{yx}[k]$ se puede extender a valores de rezago negativo. En nuestro ejemplo $c_{yx}[-1\ ]$ viene a ser valor promedio de los productos IOS[Dic] PREC[Ene], IOS[Ene] PREC[Feb], IOS[Feb] PREC[Mar], etc. Pero ese valor medio coincide con $c_{yx}[1\ ]$. En general, $c_{yx}[-k\ ] = c_{xy}[k\ ]$.

f) El valor en rezago *0* es la varianza cruzada entre las variables *x* y *y*.

g) Cuando *x* = *y*, la correlación cruzada coincide con la función de autocorrelación.

3.1.1 Ensayo de significación.

En el caso que *x* y *y* son series aleatorias no correlacionadas, la varianza de los valores estimados de la función de correlación cruzada puede ser estimada como

$$\sigma^2 = \frac{1}{N}\sum_{m=-P}^{+P} \rho_X[m]\rho_Y[m] \qquad .$$

ρ_X y ρ_Y son las funciones de autocorrelación de *x* y *y*, y *N* el número nominal de grados de libertad. σ se conoce como el error estándar de rezago grande ("large lag standard error"). Los niveles de significación son entonces 1.645σ , 2.0σ y 2.58σ a los niveles de confianza de 90, 95 y 99 %, respectivamente. Si los valores estimados de correlación cruzada exceden los niveles de significación, entonces se rechaza la hipótesis nula (ie, las series no están correlacionadas) y se acepta la hipótesis alternativa (las series están correlacionadas). El programa `wfcc.exe` calcula los niveles de confianza al 90 y 95%.

3.1.2 Consideraciones prácticas

La experiencia de los foros climáticos regionales ha sido que correlacionar secuencias mensuales no da generalmente resultados satisfactorios porque se mezclan comportamientos estacionales diferentes (por ejemplo, estación seca con la lluviosa, inicio de la estación lluviosa con el final, etc.) que tienden a disminuir las correlaciones y a dificultar las interpretaciones. Para los efectos de los foros es más productivo correlacionar valores promedio de varios meses (usualmente tres) de cada año, convirtiendo en efecto las secuencias mensuales en anuales. Por ejemplo, se puede

correlacionar el promedio de la TSM del ATN para los meses de junio, julio y agosto con un índice relacionado con la precipitación de un conjunto de estaciones de Centroamérica durante los meses de diciembre, enero y febrero (EOF-PCP). La Fig. 3.1.2.1 muestra la función de correlación cruzada entre estas dos secuencias para los años 1958 a 1999 y rezagos de -5 a 5. Las líneas rojas y las verdes definen los niveles de significación al 90 y 95%, respectivamente. Nótese que hay dos valores significativos al 95%: un valor positivo en el rezago +1 (EOF-PCP adelantado un año respecto al ATN), y un valor negativo en rezago 0 (las dos secuencias sin desplazamiento). El resto de valores no son significativos al 90%.

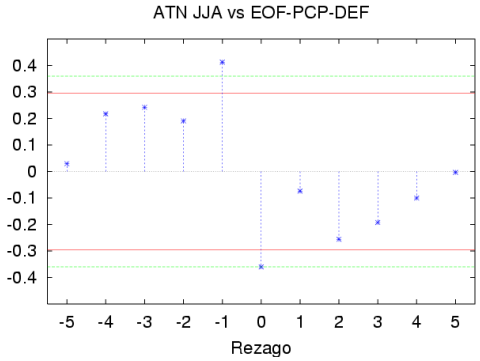


Fig. 3.1.2.1 Función de correlación cruzada entre el promedio de la temperatura superficial del Atlántico Tropical Norte (ATN) para los meses de junio, julio y agosto con un índice relacionado con la precipitación de un conjunto de estaciones de Centroamérica durante los meses de diciembre, enero y febrero (EOF-PCP)[5].

Para los efectos del RCOF no tiene sentido buscar correlaciones con rezagos mayores a un año y por lo tanto el programa `wfcc.exe` sólo calcula la función de correlación cruzada para los rezagos $k = 0, \pm 1$.

3.2 Construcción e interpretación de la tabla de contingencia

La asociación entre dos variables continuas está determinada por la distribución de probabilidad conjunta (como la probabilidad condicional para la precipitación dado que el índice Niño 3 se encuentre en su tercil más alto). En general esta distribución es desconocida, y en su defecto se sustituye por la "tabla de contingencia", derivada de las muestras tomadas de las variables poblacionales. En el caso de las aplicaciones a los RCOFs, las muestras son las series de tiempo de los valores mensuales en las estaciones meteorológicas y los índices climáticos derivados de variables predictoras (Niño 3, ATN, MEI, etc.). Para construir esta tabla las variables continuas se dividen en categorías, digamos la primera variable o variable independiente X en M categorías y la segunda variable o variable dependiente Y, en N categorías. Cada pareja de valores (x_i, y_i) pertenece a una y sólo una de las M x N categorías conjuntas (i-ésima y j-ésima). Luego

[5] Ver: Alfaro, E., 2002: Some Characteristics of the Annual Precipitation Cycle in Central America and their Relationships with its Surrounding Tropical Oceans. *Tópicos Meteorológicos y Oceanográficos*, 9(2), 88-103.

se calculan las frecuencias empíricas f_{ij} que son el número de parejas que pertenecen a la categoría *ij*. Si la asociación entre las dos variables es muy débil, la población de la *M* x *N* categorías conjuntas es similar, debido a que ninguna es significativamente más probable que las otras y hay muy poco argumento para la predicción. En un gráfico de tres dimensiones la superficie aparece plana porque los valores de *X* dentro de la categoría *i* pueden estar asociados indistintamente con valores que pertenecen a cualquiera de las *N* categorías de *Y* (Fig. 3.2.1.a). Si la asociación lineal es fuerte, la superficie alcanza valores grandes a lo largo de una de las diagonales y valores bajos en las esquinas que no pertenecen a la diagonal (Fig. 3.2.1.b). En el caso de asociación positiva los valores altos se dan a lo largo de la diagonal mayor y en el caso de asociación negativa, a lo largo de la diagonal menor.

a) b)

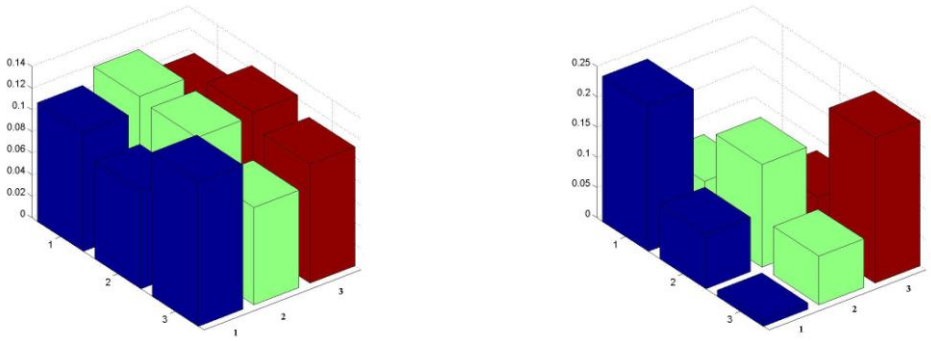


Figura 3.2.1. Probabilidades empíricas para: a) un caso en que la asociación entre las variables es débil, nótese que la frecuencia en cada una de las casillas es muy similar y b) un caso en que la asociación entre las variables es fuerte, nótese que las frecuencias a lo largo de la diagonal son superiores al resto de las categorías.

Al dividir las frecuencias empíricas por el número total de parejas se obtienen las probabilidades empíricas. Es decir $p_{ij} = f_{ij} / n$ es la probabilidad empírica que una pareja de valores pertenezca a la categoría conjunta *ij*. En los RCOFs los valores individuales de p_{ij} o f_{ij} pueden ser expresados en términos de porcentajes, P_{ij}, en función de los *M* escenarios de la variable independiente o predictor, es decir,

$$P_{ij} = \frac{p_{ij}}{\sum_{j=1}^{N} p_{ij}} \times 100 \text{ , o,} \qquad P_{ij} = \frac{f_{ij}}{\sum_{j=1}^{N} f_{ij}} \times 100 .$$

Se han definido estadísticos que cuantifican el grado de asociación de una manera objetiva y que serán discutidos en la sección 3.4.

3.3 La tabla de contingencia *3* x *3*

A primera vista pareciera conveniente usar un número grande de categorías para alcanzar una mayor resolución. En la práctica se ha encontrado que un número alto de

categorías es difícil de interpretar por el gran número de posibilidades a considerar. Además, se ocuparía un número muy grande de datos para lograr un análisis estable de las *M* x *N* categorías. Al dividir las dos variables en terciles, es decir, haciendo *M* = *N* = *3*, se obtienen *9* categorías conjuntas que permiten cierto grado de resolución y un número manejable de posibilidades. Por otro lado, una manera pragmática de enfocar la selección de *3* x *3* es que frecuentemente se busca un pronóstico que sea entendido fácilmente por el público en general usando conceptos tales como: *normal, arriba de lo normal* y *abajo de lo normal*.

Supongamos que se quiere hacer una validación cruzada entre un conjunto de observaciones (variable independiente, o predictora) y un conjunto de valores cuya proyección a futuro se desea pronosticar (variable dependiente, predictante). Como ejemplo podríamos tener como variable predictora algún índice de El Niño-Oscilación del Sur o ENOS (p.e. IOS), mientras que la variable a pronosticar podría ser la lluvia esperada en alguna región climática o cuenca del país, medido normalmente por el promedio de un conjunto de estaciones pluviométricas. Las observaciones se dividen en terciles: valores bajos (Obs. Bajo), valores normales (Obs. Normal) y valores altos (Obs. Alto). Cada observación (es decir renglón o fila del archivo de datos) tiene que pertenecer a una sola categoría (las categorías son excluyentes) y las tres categorías agotan todas las posibilidades dentro de la "muestra" o conjunto de observaciones tomadas (las categorías son exhaustivas). Esto se ilustra en el diagrama de Venn de la Fig. 3.3.1. La Fig. 3.3.2 muestra el diagrama de Venn para los pronósticos. Como las categorías son terciles, cada categoría es un tercio del área total. Al considerar las parejas de (valores observado, valor pronosticado) el diagrama de Venn para las parejas toma la forma de la Fig. 3.3.3 en la cual se identifican nueve categorías excluyentes y exhaustivas para las parejas. Por ejemplo, un valor bajo de la observación puede estar asociado con un valor bajo (OB-PB), uno normal (OB-PN) o uno alto (OB-PA) del valor pronosticado.

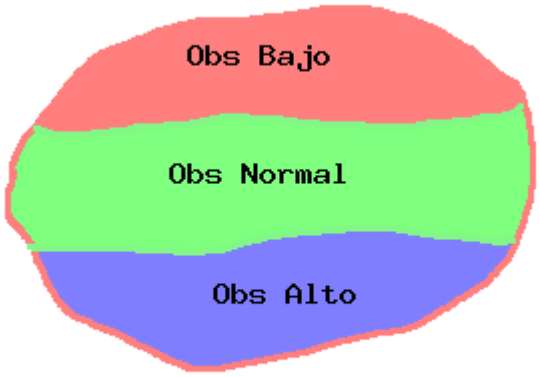


Figura 3.3.1. Diagrama de Venn de la variable independiente categorizada en terciles.

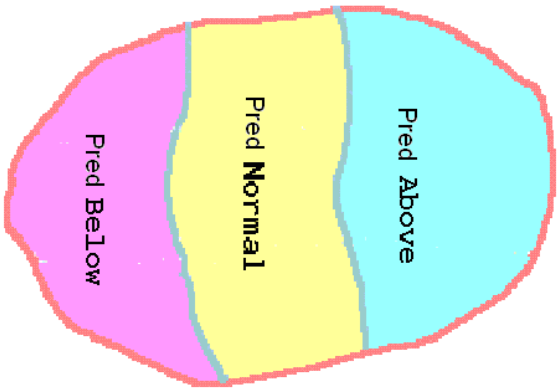


Figura 3.3.2. Diagrama de Venn de la variable dependiente categorizada en terciles.

En otras palabras, hemos determinado para cada una de las dos variables, cuales son las categorías históricas que se consideran normal, arriba de lo normal o debajo de lo normal, en ausencia de cualquier otro criterio. Pero si sospechamos que la categoría de la variable predictora afecta la distribución de lluvias (p.e., más lluvias con La Niña, menos lluvias con El Niño), podemos comprobar esto o rechazarlo repartiendo los datos entre las categorías conjuntas.

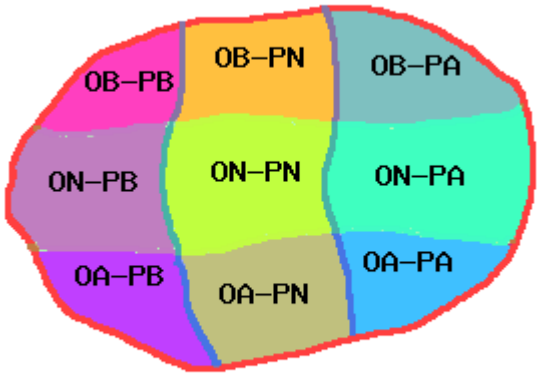


Figura 3.3.3. Diagrama de Venn de las categorías conjuntas.

La Fig. 3.3.3 muestra que el área de cada una de las nueve posibilidades es aproximadamente del mismo valor. Este es el caso cuando la asociación entre las observaciones y el pronóstico es débil o aleatorio (similar a lo mostrado en la Fig. 3.2.1.a). En este caso la condición de la variable predictora no rinde un efecto significativo sobre la variable a pronosticar. En otras palabras y refiriéndonos al ejemplo anterior, al estar el Océano Pacífico en estado normal, o Niño o Niña, la distribución de lluvias no se aparta de manera significativa de lo esperado por la climatología (Fig. 3.3.2): [1/3, 1/3, 1/3]. Este caso teórico de independencia estadística entre la observación y el pronóstico se utiliza como punto de referencia para evaluar la bondad de un pronóstico. Mientras más difieran las propiedades estadísticas de una tabla de

contingencia respecto a las propiedades de la tabla de contingencia de variables independientes (Fig. 3.3.3), se considera más fuerte la asociación entre la variable independiente y dependiente. Muchos de los estadísticos y de los ensayos de significación utilizados cuantifican que tan diferente es una tabla de contingencia dada con respecto a la tabla de contingencia del caso de variables independientes.

Para eventos independientes, la probabilidad que ocurran simultáneamente es multiplicativa. Por lo tanto, la probabilidad que una pareja (observación, pronóstico) corresponda al caso ON-PA, por ejemplo es

$$\Pr\{O_2 \cap P_3\} = \Pr\{O_2\} \Pr\{P_3\} = \frac{1}{3}\frac{1}{3} = \frac{1}{9}.$$

En general,

$$\Pr\{O_i \cap P_j\} = \Pr\{O_i\} \Pr\{P_j\} = \frac{1}{3}\frac{1}{3} = \frac{1}{9}.$$

En pocas palabras, si son independientes, una observación (predictor) dada puede estar asociada indistintamente con valores bajos, normales o altos de la variable de pronóstico (predictante).

El pronóstico es exitoso si una observación baja (normal, alta) coincide con valor de pronóstico bajo (normal, alto). Técnicamente:

$$\Pr\{Exito\} = \Pr\{(O_1 \cap P_1) \cup (O_2 \cap P_2) \cup (O_3 \cap P_3\} = \Pr\{O_1 \cap P_1\} + \Pr\{O_2 \cap P_2\} + \Pr\{O_3 \cap P_3\} = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}.$$

Nótese que la probabilidad de éxito corresponde al área total a lo largo de la diagonal mayor.

En el caso de la correlación negativa, el pronóstico sería exitoso si una observación baja (normal, alta) coincide con valor de pronóstico alto (normal, bajo), es decir,

$$\Pr\{Exito\} = \Pr\{(O_1 \cap P_3) \cup (O_2 \cap P_2) \cup (O_3 \cap P_1\} = \Pr\{O_1 \cap P_3\} + \Pr\{O_2 \cap P_2\} + \Pr\{O_3 \cap P_1\} = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}.$$

Ahora la probabilidad de éxito corresponde al área a lo largo de la diagonal menor.

3.4 Análisis Estadístico[6]

3.4.1 Estadísticos del análisis

- La prueba chi cuadrado, χ^2

Otra medida de discrepancia entre las frecuencias observadas (f_{ij}) y esperadas (e_{ij}) viene proporcionada por el estadístico χ^2 dado por

$$\chi^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}},$$ donde la frecuencia total es *n*.

El estadístico chi cuadrado evalúa que tanto se aleja una tabla de contingencia del caso aleatorio y se puede aplicar a tablas de contingencia rectangulares. En nuestro caso es cuadrada de 3 x 3. El número de grados de libertad, en una tabla de entradas múltiples, si las filas y columnas son variables independientes, es (M-1) x (N-1). El número de grados de libertad en nuestro caso corresponde a (3-1) x (3-1) = 4, ya que las frecuencias esperadas se calculan sin recurrir a estimaciones muestrales de los parámetros de la población. Si la tabla de contingencia 3 x 3 es aleatoria, la suma de cualquier fila o columna es n/3, y por la tanto para cualquier valor de (i,j), eij = (n/3) (n/3)/ (n) = n /9. La significación del estadístico χ^2 se determina sobre la base de la hipótesis H_0. Si bajo tal hipótesis el valor calculado para χ^2 es mayor que algún valor crítico (tal como $\chi^2_{.95}$ o $\chi^2_{.99}$, que son los valores de significación 0.05 y 0.01 respectivamente), se debe concluir que las frecuencias observadas difieren significativamente de las frecuencias esperadas y se rechaza H_0 al correspondiente nivel de significación. La significación del estadístico χ^2 se determina por la probabilidad acumulada empezando de 0 al valor χ^2. Valores independientes de filas y columnas tienen valores de significación que tienden a 0. Cuando las variables independiente y dependiente están fuertemente asociadas positivamente, χ^2 obtiene valores altos y la significación tiende a 1 en su extremo máximo posible. En términos más gráficos, la Figura 3.2.1(a) tiene una valor de χ^2; más bajo que la Figura 3.2.1(b).

3.4.3 Coeficiente de correlación de Pearson

Este coeficiente es apropiado cuando ambas variables se han categorizado y tiene un rango de $-1 \leq r \leq 1$, este coeficiente es calculado utilizando las siguientes relaciones,

[6] Se incluyen algunas sugerencias dadas por el Dr. Luis Cid del Departamento de Estadística de la Universidad de Concepción, Chile.

$$r = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}$$

donde,

$$ss_r = \sum_i \sum_j f_{ij} (R_i - \overline{R})^2 ,$$

$$ss_c = \sum_i \sum_j f_{ij} (C_i - \overline{C})^2$$

y

$$ss_{rc} = \sum_i \sum_j f_{ij} (R_i - \overline{R})(C_i - \overline{C}) .$$

R_i y C_j son respectivamente los totales de la fila i y columna j.

El ensayo de significación para el coeficiente de correlación Pearson utiliza el estadístico normalizado r^* que posee asintóticamente una distribución normal bajo la hipótesis nula. Este estadístico se define como

$$r^* = \frac{r}{\sqrt{\text{var}_0(r)}},$$

donde

$$\text{var}_0(r) = \frac{\sum_i \sum_j f_{ij} (R_i - \overline{R})^2 (C_j - \overline{C})^2 - ss_{rc}^2 / n}{ss_r \, ss_c}.$$

Esta es la varianza asintótica derivada del muestreo multinomial dentro del marco de una tabla de contingencia. Ella difiere de la forma obtenida bajo la suposición de que las dos variables son continuas y distribuidas normalmente[7].El valor de significación corresponde a la probabilidad acumulada a la derecha de la distribución normal. De esta manera cuando, **r*** tiene valores cercanos a cero (las variables no están correlacionadas), la significación corresponde a 0.5. La significación de variables positivamente (negativamente) correlacionadas tiende a 1 (0).

3.4.2 Nota sobre la Validación Cruzada[8]

[7] Brown, M.B. and Benedetti, J.K. 1977. Sampling behavior of tests for correlation in two way contingency tables. J. Am. Stat. Ass., 72, 309-315.

[8] Ver: - Ward, N., and C. Folland, 1991. Prediction of seasonal rainfall in the North Nordeste of Brazil using eigenvectors of Sea Surface Temperature. *Int. J. of Climatol.*, **11**, 711-743.

- Wilks, D., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press. 465 pp.

La validación cruzada es una técnica de remuestreo que opera en forma similar a las pruebas de bootstrap y permutación, dividiendo repetidamente el conjunto total de datos en un subconjunto de control y otro de verificación. Lo más común es que el primero sea de tamaño *n-1* y el segundo de *1*, haciendo *n* particiones diferentes de los datos. La forma de operar de esta técnica es la siguiente[9]:

1) Se remueve el primer dato de la variable dependiente y el o los primeros datos de las variables independientes y se recalcula el modelo con los *n-1* pares de datos restantes o de control, a este modelo se le llama modelo reducido.
2) Usando el modelo reducido calculado en 1), se estima el valor removido de la variable dependiente pero usando el o los datos removidos de las variables independientes. Este valor estimado, para fines de comparación, se divide entre el coeficiente de determinación múltiple, R, del modelo reducido. A estos valores estimados se les llama estimados inflados. A modo de ejemplo y para el caso de nuestro enfoque, diríamos entonces que si la tabla de contingencia (de los *n-1* valores restantes) y la categoría de la observación independiente #1 removida predicen que la observación dependiente #1 removida estuviera en su categoría *abajo de lo normal* con mayor probabilidad que en las otras categorías (*normal o arriba de lo normal*), entonces se le asigna un pronóstico a la categoría *abajo de lo normal*.
3) Se reincorpora el dato removido de las variables dependiente e independientes al conjunto total de datos y se remueven los siguientes (observación #2), repitiéndose el proceso de los puntos 1) y 2) en forma sucesiva hasta tener *n* modelos reducidos y *n* datos estimados, donde *n* es la longitud de las series de tiempo.

- Porcentaje de Falsas Alarmas y de Detección: Abajo y Arriba de lo Normal.

El evento de falsa alarma ocurre cuando un pronóstico no se materializa (o sea, falla). En nuestro ejemplo, comprende la situación donde el valor dependiente removido resulte *normal* o *superior a lo normal* en circunstancias que el valor independiente removido hubiera indicado un pronóstico *abajo de lo normal*. De la misma manera se califica cada una de las *n-1* pruebas sucesivas como *éxito* o *falsa alarma*. Donde los estadísticos de diagnóstico dados por el programa corresponden a las proporciones de *éxito* o *falsa alarma* (entre las *n* pruebas) *debajo de lo normal* y *arriba de lo normal*. Luego, estas proporciones se comparan a las esperadas para el caso de variables independientes.

Si usamos la tabla de contingencia como referencia, calculamos los valores de *Falsas Alarmas* y *Porcentajes de Detección Abajo* y *Arriba de lo Normal* como:

$FARBN = f_{13} / (f_{13} + f_{23} + f_{33})$, $FARAN = f_{31} / (f_{11} + f_{21} + f_{31})$,

$PODBN = f_{11} / (f_{11} + f_{12} + f_{13})$, y $PODAN = f_{33} / (f_{31} + f_{32} + f_{33})$, respectivamente.

[9] La función crosval.m incuida en el directorio programasfuente realiza este calculo.

A continuación profundizaremos en estos conceptos para el lector interesado en los detalles matemáticos, de otra manera el usuario puede continuar a la descripción del siguiente estadístico sin que esto afecte la comprensión de los mismos.

La probabilidad de falsas alarmas asociadas a un valor bajo de observación se define:

$$\Pr\{\ \textit{Falsa alarma} \mid O_1\} = \Pr\{P_2 \cup P_3 \mid O_1\} = \Pr\{P_2 \mid O_1\} + \Pr\{P_3 \mid O_1\} = \frac{\Pr\{P_2 \cap O_1\}}{\Pr\{O_1\}} + \frac{\Pr\{P_3 \cap O_1\}}{\Pr\{O_1\}} \quad .$$

Si hay independencia:

$$\Pr\{\textit{Falsa alarma} \mid O_1\} = \frac{1/9}{1/3} + \frac{1/9}{1/3} = \frac{2}{3}.$$

El concepto se puede simplificar considerando los valores de las esquinas de la tabla de contingencia fuera de la diagonal mayor. Así, la probabilidad de falsa alarma debajo de lo normal, o *FARBN*, y la probabilidad de falsa alarma arriba de lo normal, o *FARAN*, se definen como:

$$FARBN = \Pr\{\ P_3 \mid O_1\} = \frac{\Pr\{P_3 \cap O_1\}}{\Pr\{O_1\}}, \ y \quad FARAN = \Pr\{\ P_1 \mid O_3\} = \frac{\Pr\{P_1 \cap O_3\}}{\Pr\{O_3\}},$$

respectivamente.

Con los valores de las esquinas a lo largo de la diagonal mayor podemos definir probabilidades de detección arriba y abajo de lo normal, *PODAN* y *PODBN*, respectivamente. De manera análoga,

$$PODBN = \Pr\{\ P_1 \mid O_1\} = \frac{\Pr\{P_1 \cap O_1\}}{\Pr\{O_1\}}, \ y \quad PODAN = \Pr\{\ P_3 \mid O_3\} = \frac{\Pr\{P_3 \cap O_3\}}{\Pr\{O_3\}}.$$

Si hay independencia *FARBN = FARAN = PODBN = PODAN =1/3*. Una asociación fuerte entre la variable independiente y dependiente disminuye los valores de *FARBN* y *FARAN* y aumenta los de *PODBN* y *PODAN*.

- Razón de Aciertos (*Hit Rate*)

Como se mencionó anteriormente, la probabilidad de éxito corresponde al área a lo largo de la diagonal mayor o menor según sea el caso. La probabilidad de éxito en

porcentaje se conoce como razón de aciertos (*HR*, *hit rate* en inglés). Este se calcula como:

$HR = (f_{11} + f_{22} + f_{33}) / n * 100$, para una correlación positiva y

donde ya se vió que por azar, esperamos un acierto $C = 1/3$, o sea, un 33.33%.

- Puntaje de Destreza (*Skill Score*)

Podemos transformar la información que nos otorga el coeficiente *HR*, para construir el puntaje de destreza o *SS* (*Skill Score*, por sus siglas en inglés), como:

$$SS = \frac{HR - 33.33}{100 - 33.33} \times 100.$$

Nótese que cuando el pronóstico es totalmente al azar, $SS = 0$, y cuando la correspondencia es exacta ($\Pr\{Éxito\} = 1$) $SS = 100$, o sea un conjunto perfecto de aciertos. En este último caso se nos está diciendo que las probabilidades empíricas fuera de la diagonal (mayor o menor, según el caso), se anulan. Valores negativos del *SS*, nos indicarían que los desaciertos (*misses*, en inglés) dominan en nuestro análisis.

- Error Lineal en el Espacio de Probabilidades (*LEPS Score*)

Otro coeficiente útil es el coeficiente del error lineal en el espacio de probabilidades o *LEPS score* por sus siglas en inglés. Este coeficiente es similar al *SS* excepto que ahora los pronósticos que tienen dos terciles de error son castigados más fuertemente que aquellos que tienen solo uno y lo podemos expresar como:

$LEPS = (z_1 / z_2) * 100$,

donde z_1 es la sumatoria de las frecuencias ponderadas, para la correlación positiva lo tomaríamos como:

$$z_1 = 1.35 * f_{11} - 0.15 * f_{12} - 1.20 * f_{13} - 0.15 * f_{21} + 0.30 * f_{22} - 0.15 * f_{23} - 1.20 * f_{31} - 0.15 * f_{32} + 1.35 * f_{33},$$

ahora, z_2 es la sumatoria de las frecuencias ponderadas de un conjunto perfecto de aciertos, es decir *n* en nuestro caso. Nótese que si el pronóstico fuera al azar y todas las frecuencias empíricas de la tabla de contingencia tendieran al mismo valor, z_1 y por lo tanto *LEPS* tendería a 0. Por otro lado si tuviéramos un pronóstico perfecto la razón z_1/z_2 tiende a 1 y el *LEPS* tendería a *100*. Al igual que en el *SS*, valores negativos del *LEPS*, nos indicarían que los desaciertos dominan nuestro análisis.

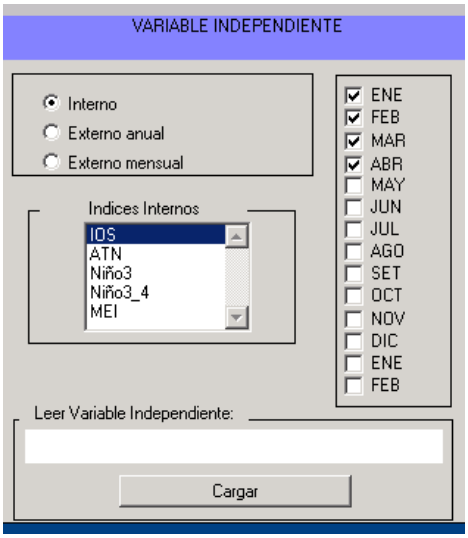
4. Ejemplos, Aplicaciones a los foros climáticos

4.1 Aplicación a los foros climáticos I

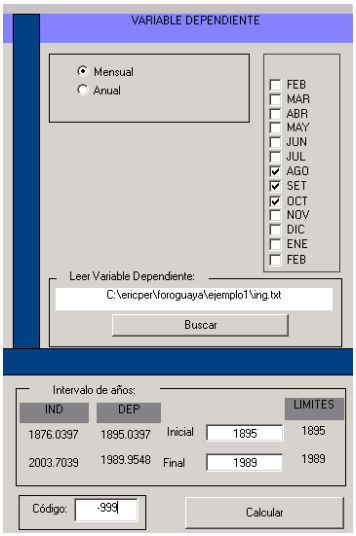
Para este primer ejemplo usaremos los datos de precipitación de la estación Ingenio San Antonio (87.05° W, 12.53° N, 35 msnm), ubicada en la costa pacífica de Nicaragua, con registros mensuales de 1895 a 1989. Los datos se encuentran en el directorio ...\ejemplo1\ing.txt, y el código para los datos faltantes es –999. El archivo ing.txt contiene dos columnas, la primera es el año mas la fracción del mes (vector de tiempo) y la segunda el acumulado de precipitación mensual en mm para la estación del Ing. San Antonio:

```
% tiempo      Precip.-Ingenio San Antonio
1895.039726    0.000000
1895.123058    0.000000
1895.206390    0.000000
1895.289722    0.000000
    ⋮            ⋮
1989.704799  664.000000
1989.788131  252.000000
1989.871463   41.000000
1989.954795    4.000000
```

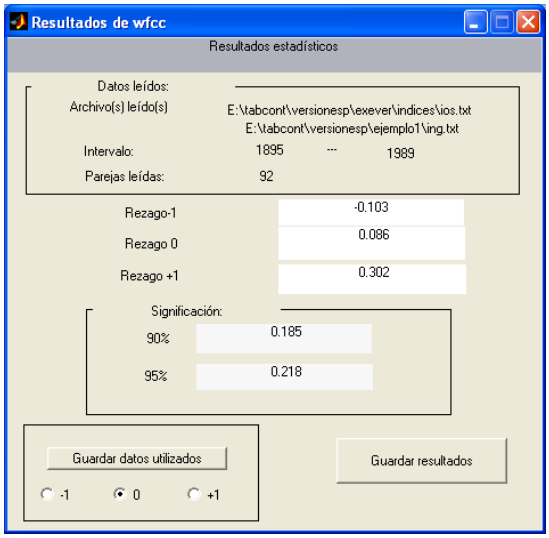
Exploraremos en este ejemplo la relación entre la precipitación de ASO en el Ing. San Antonio y el Índice de Oscilación del Sur o IOS para los periodos EFMA, MJJ, ASO, ND. Como primer paso al ejecutar el programa wfcc.exe, debemos marcar el círculo Interno bajo VARIABLE INDEPENDIENTE, ya que el IOS es uno de los índices climáticos que se brindan con el programa, luego en los meses marcamos ENE, FEB, MAR y ABR, posteriormente hacemos clic en cargar.



Seguidamente, bajo VARIABLE DEPENDIENTE, marcamos Mensual, ya que el archivo ing.txt contiene datos mensuales, luego en los meses marcamos AGO, SET y OCT, posteriormente buscamos el archivo ing.txt en el directorio correspondiente y hacemos clic en abrir u open. Luego de esto elegimos el periodo a usar, para nuestro ejemplo 1895 en Inicial y 1989 en Final, también debemos digitar el Código del dato faltante, -999.



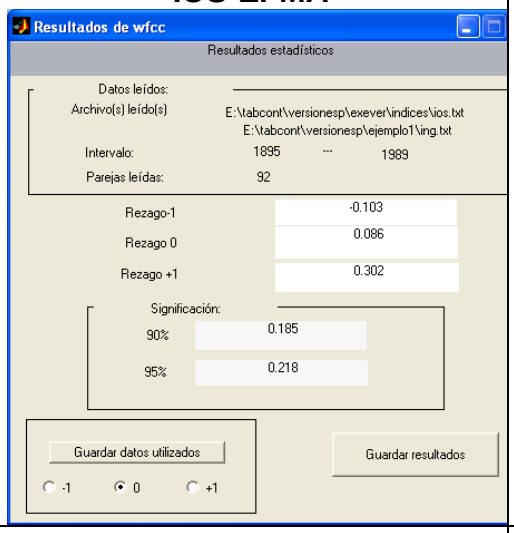
Una vez hecho lo anterior hacemos clic en Calcular, lo que despliega la siguiente pantalla:



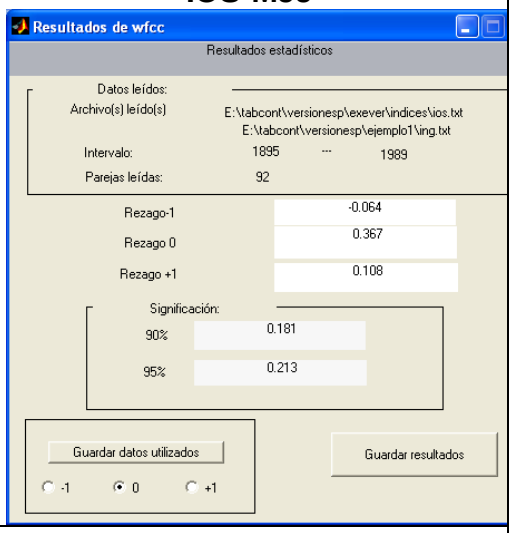
Estos resultados se refieren a la función de correlación cruzada o FCC entre el IOS-EFMA del año anterior a la precipitación en el Ing. San Antonio durante ASO o rezago –1, entre el IOS-EFMA del mismo año a la precipitación en el Ing. San Antonio durante ASO o rezago 0 y entre el IOS-EFMA del año siguiente a la precipitación en el Ing. San Antonio durante ASO o rezago +1. También se incluyen los niveles de significación de

estos valores de estas correlaciones al 90 y 95%. Repitiendo los pasos anteriores pero seleccionando como periodos en la variable independiente MJJ, ASO y ND obtenemos los siguientes resultados:

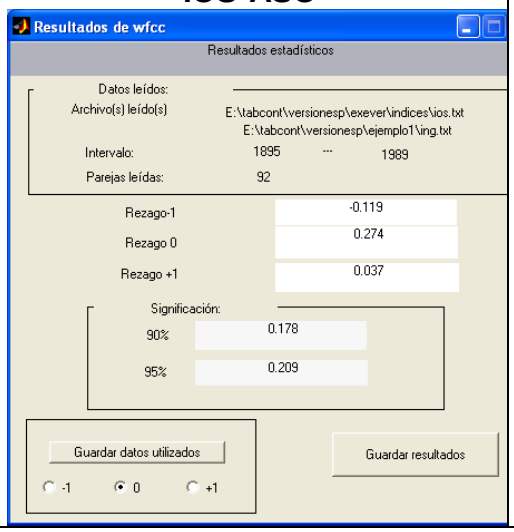
IOS-EFMA



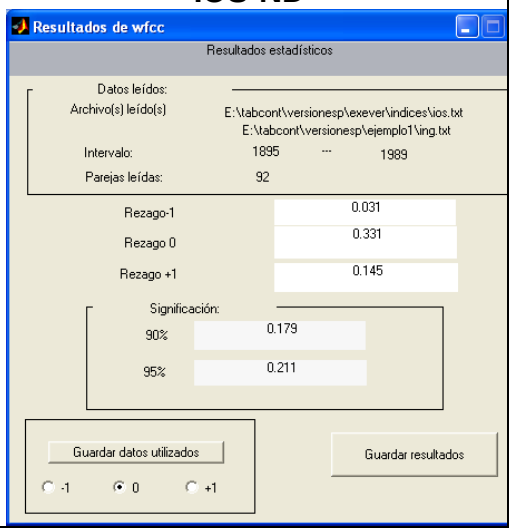
IOS-MJJ



IOS-ASO



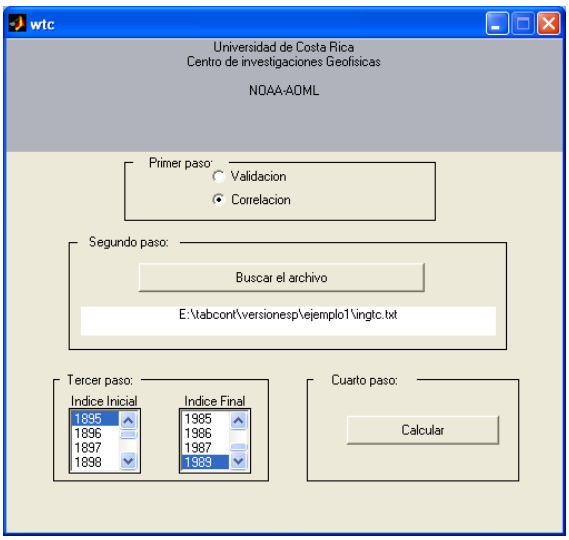
IOS-ND



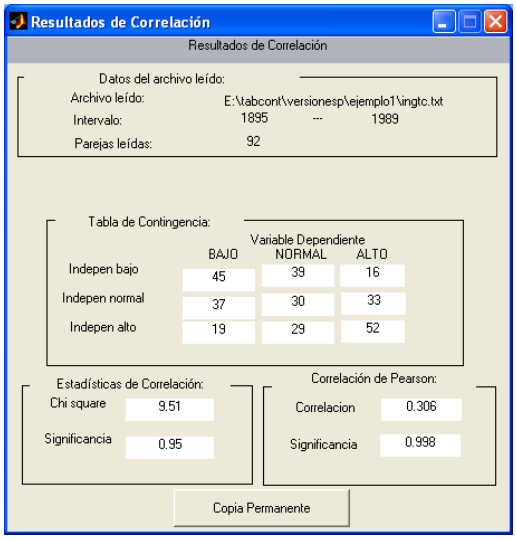
Nótese que el valor de la FCC entre el IOS-MJJ y la precipitación de ASO del año 0 es 0.367, significativo al 95%, por lo que en principio lo podriamos usar en un esquema clásico de predicción. Para proceder de esta forma hacemos `clic` en el botón `Guardar datos utilizados` en la pantalla de los resultados del IOS-MJJ y los salvamos en el archivo `...\ejemplo\ingios.txt`. El archivo `ingios.txt` contiene tres columnas, la primera es el año, la segunda el promedio de los valores del IOS para el trimestre MJJ y la tercera el promedio las anomalías de la precipitación en la estación del Ing. San Antonio para el trimestre ASO. Las líneas precedidas por el símbolo % se toman como comentarios. Como no se pueden incluir en el análisis contingente los datos faltantes, procedemos a anteponer entonces el símbolo % delante de las líneas que contengan -999 en alguna de sus columnas, utilizando cualquier procesador de texto conveniente:

```
% Archivo de series anuales:
% Compuesto de los siguientes archivos fuente:
% Archivo leído C:\ericper\viernes10oct\exever\indices\ios.txt
% Archivo leído C:\ericper\foroguaya\ejemplo1\ing.txt
% Año inicial: 1895.00
% Año final: 1989.00
% Año            IOS-MJJ(0)      Pre. Ing. San Antonio-ASO
  1.8950e+003  -4.4333e+000       4.1033e+002
  1.8960e+003  -3.1133e+001       2.4833e+002
  1.8970e+003  -6.3333e+000       4.7300e+002
  1.8980e+003   6.3333e-001       3.3133e+002
  1.8990e+003  -7.8000e+000       3.5400e+002
  1.9000e+003   9.5667e+000       3.8067e+002
% 1.9010e+003   1.1300e+001      -9.9900e+002
% 1.9020e+003   3.9333e+000      -9.9900e+002
  1.9030e+003   4.3667e+000       3.2567e+002
```

Luego de salvar este archivo ejecutamos el programa `wtc.exe`. En `Primer paso` selecionamos el círculo `Correlacion`, en `Segundo paso`, buscamos el archivo `ingios.txt`, en `Tercer paso` selecionamos 1895 en `Indice inicial` y 1989 en `Indice Final`.



Por último en `Cuarto paso` hacemos clic en `Calcular`, lo que despliega la siguiente pantalla:



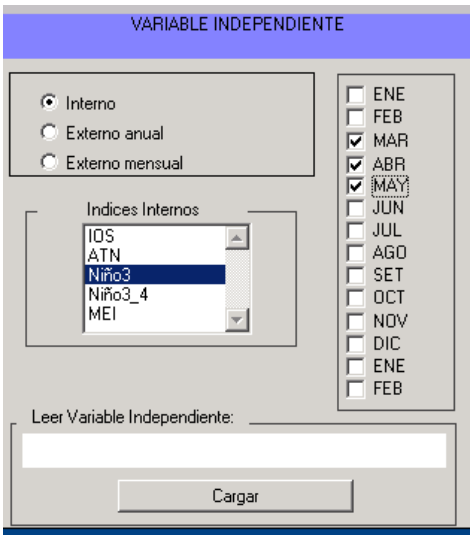
Como parte del ejemplo, si suponemos que observamos un valor negativo del IOS durante MJJ ubicado en el primer tercil, nuestra sugerencia para el foro climático sobre las probabilidades esperadas en la precipitación del Ing. San Antonio durante ASO será: 45% BN, 39% DN y 16% AN.

4.2 Aplicación a los foros climáticos II

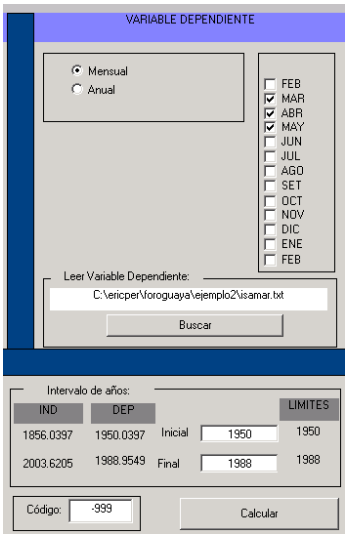
Para este segundo ejemplo usaremos los datos de precipitación de la estación Isabel María (79.56° W, 1.83° N, 4 msnm), del Ecuador, con registros mensuales de 1950 a 1988. Los datos se encuentran en el directorio ...\ejemplo2\isamar.txt, y el código para los datos faltantes es –999. El archivo isamar.txt contiene dos columnas, la primera es el año mas la fracción del mes (vector de tiempo) y la segunda el acumulado de precipitación mensual en mm para la estación de Isabel María:

```
% tiempo           Precip.-Isabel Maria
  1.9500397e+003  3.6910000e+002
  1.9501231e+003  4.9310000e+002
  1.9502064e+003  3.1700000e+002
  1.9502897e+003  1.6760000e+002
    ⋮                 ⋮
  1.9887049e+003 -9.9900000e+002
  1.9887883e+003 -9.9900000e+002
  1.9888716e+003 -9.9900000e+002
  1.9889549e+003 -9.9900000e+002
```

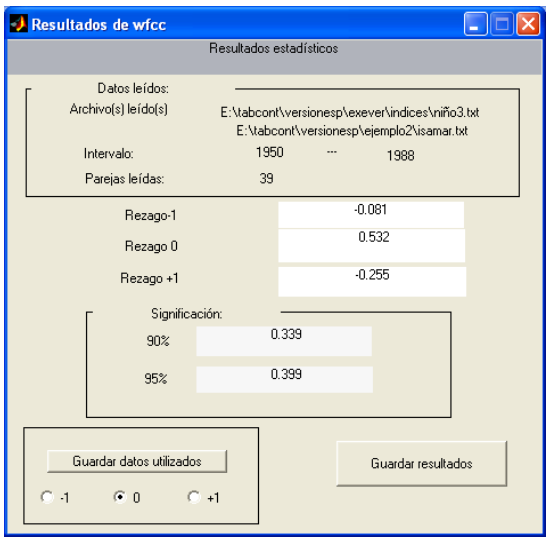
Exploraremos en este ejemplo la relación entre la precipitación de MAM en Isabel María y el Índice del Niño 3 para los periodos DEF, MAM, JJA, SON. Como primer paso al ejecutar el programa wfcc.exe, debemos marcar el círculo Interno bajo variable independiente, ya que el Niño 3 es uno de los índices climáticos que se brindan con el programa, luego en los meses marcamos MAR, ABR y MAY, posteriormente hacemos clic en cargar.



Seguidamente, bajo `variable dependiente`, marcamos `mensual`, ya que el archivo `isamar.txt` contiene datos mensuales, luego en los meses marcamos `MAR`, `ABR` y `MAY`, posteriormente buscamos el archivo `isamar.txt` en el directorio correspondiente y hacemos clic en `abrir` u `open`. Luego de esto elegimos el periodo a usar, para nuestro ejemplo 1950 en `Inicial` y 1988 en `Final`, también debemos digitar el Código del dato faltante, -999.

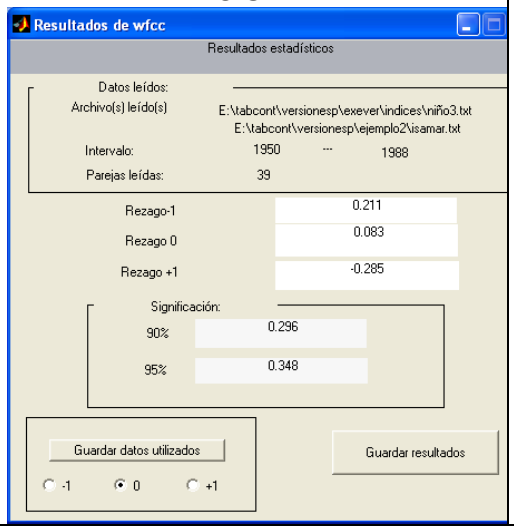


Una vez hecho lo anterior hacemos clic en `Calcular`, lo que despliega la siguiente pantalla:

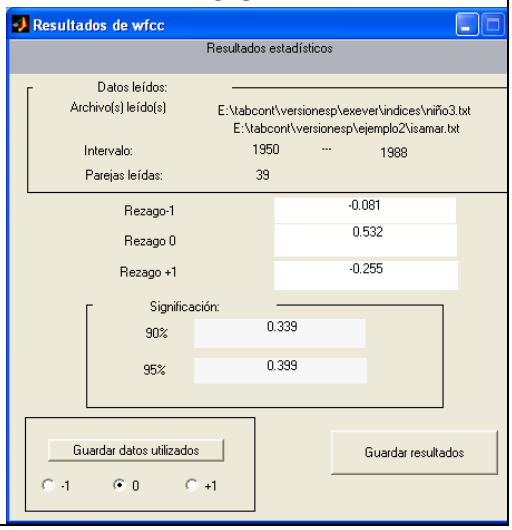


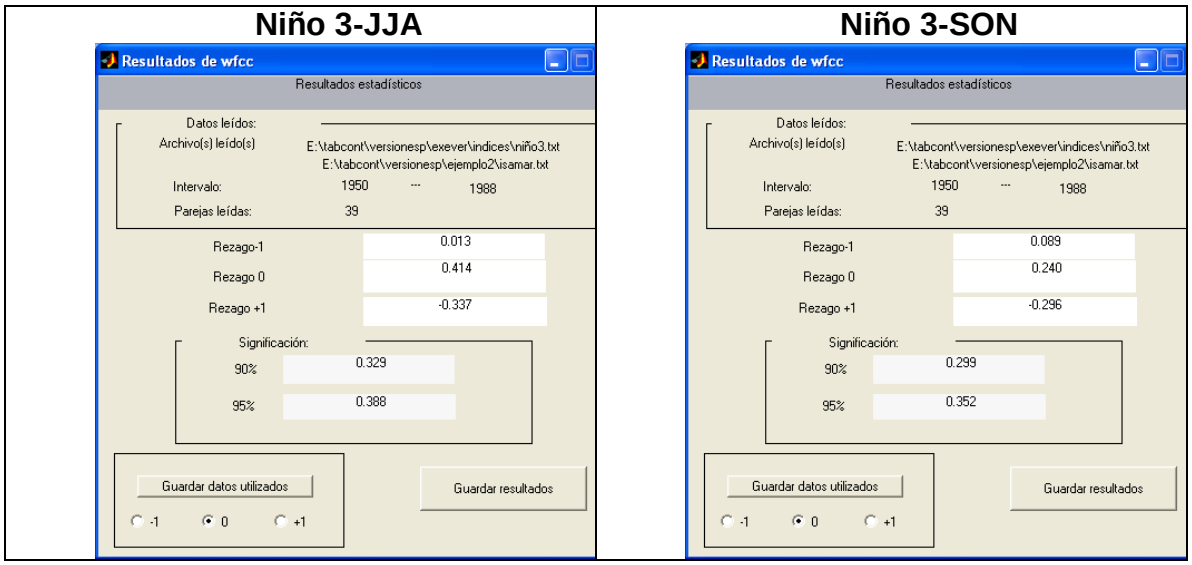
Estos resultados se refieren a la función de correlación cruzada o FCC entre Niño3-MAM del año anterior a la precipitación en Isabel María durante MAM o rezago –1, entre Niño3-MAM del mismo año a la precipitación en Isabel María durante MAM o rezago 0 y entre Niño3-MAM del año siguiente a la precipitación en Isabel María durante MAM o rezago +1. También se incluyen los niveles de significación de estos valores de estas correlaciones al 90 y 95%. Repitiendo los pasos anteriores pero seleccionando como periodos en la variable independiente DEF, JJA y SON obtenemos los siguientes resultados:

Niño 3-DEF



Niño 3-MAM



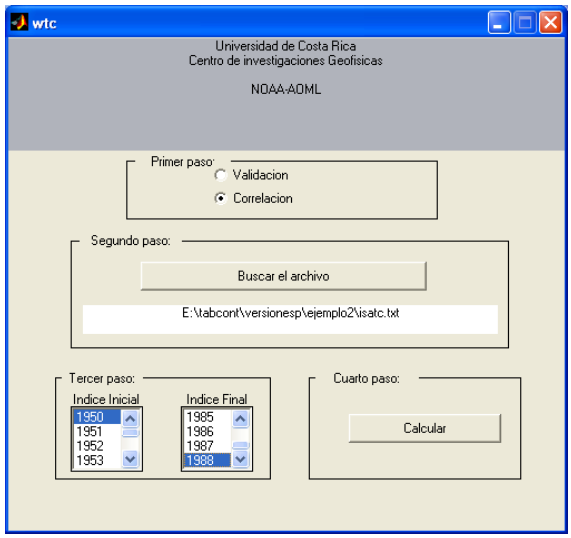


Nótese que el valor de la FCC entre el Niño 3-MAM y la precipitación de MAM del año 0 es 0.53, significativo al 95%. Para usar este resultado en predicción climática deberiamos usar el esquema conocido como "prognosis perfecta" y basarnos en los resultados de alguna predicción para escoger el escenario del índice Niño 3, ya que este es simultáneo al evento de precipitación en la estación Isabel María (rezago 0). Para proceder de esta forma hacemos `clic` en el botón `Guardar datos utilizados` en la pantalla de los resultados del Niño 3-MAM y los salvamos en el archivo `...\ejemplo2\isamarn3.txt`. El archivo `isamarn3.txt` contiene tres columnas, la primera es el año, la segunda el promedio de los valores del Niño 3 para el trimestre MAM y la tercera el promedio las anomalías de la precipitación en la estación del Isabel María para el trimestre MAM. Las líneas precedidas por el símbolo % se toman como comentarios. Como no se pueden incluir en el análisis contingente los datos faltantes, procedemos a anteponer entonces el símbolo % delante de las líneas que contengan -999 en alguna de sus columnas, utilizando cualquier procesador de texto conveniente:

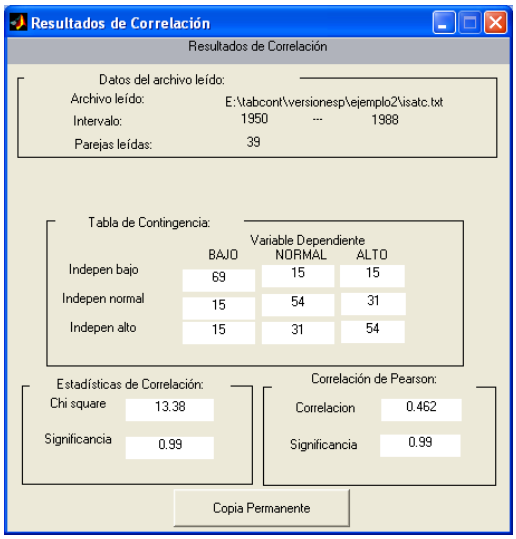
```
% Archivo de series anuales:
% Compuesto de los siguientes archivos fuente:
% Archivo leído
C:\ericper\viernes10oct\exever\indices\n3.txt
% Archivo leído C:\ericper\foroguaya\ejemplo2\isamar.txt
% Año inicial: 1950.00
% Año final: 1988.00
% Año           Niño 3-MAM(0)     Pre. Isabel Maria-MAM

  1.9500e+003   -7.7667e-001    1.6187e+002
  1.9510e+003   -8.3333e-002    2.3890e+002
  1.9520e+003   -1.3333e-001    1.2887e+002
  1.9530e+003    5.0667e-001    4.3687e+002
  1.9540e+003   -7.4000e-001    1.3697e+002
```

```
1.9550e+003   -9.3333e-001    3.0887e+002
1.9560e+003   -5.0667e-001    2.4790e+002
1.9570e+003    4.2667e-001    4.4293e+002
1.9580e+003    3.3000e-001    3.3827e+002
```

Luego de salvar este archivo ejecutamos el programa `wtc.exe`. En `Primer paso` selecionamos el círculo `Correlacion`, en `Segundo paso`, buscamos el archivo `isamarn3.txt`, en `Tercer paso` selecionamos 1950 en `Indice inicial` y 1988 en `Indice Final`:



Por último en `Cuarto paso` hacemos clic en `Calcular`, lo que despliega la siguiente pantalla:



Como parte del ejemplo, si suponemos que esperamos un valor dentro de lo normal del Niño 3 durante MAM ubicado en el segundo tercil, nuestra sugerencia para el foro

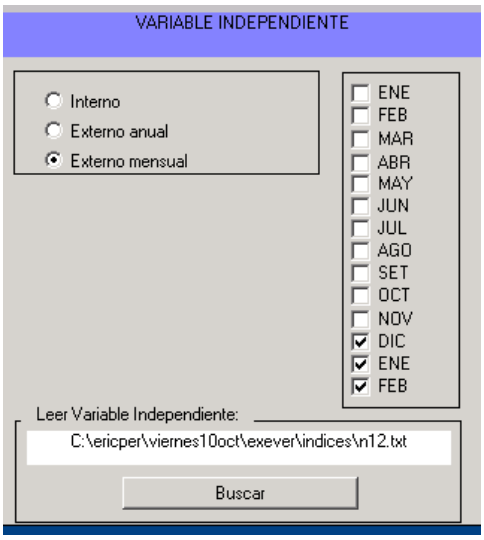
climático sobre las probabilidades esperadas en la precipitación de Isabel María durante MAM será: 15% BN, 54% DN y 31% AN.

4.3 Aplicación a los foros climáticos III

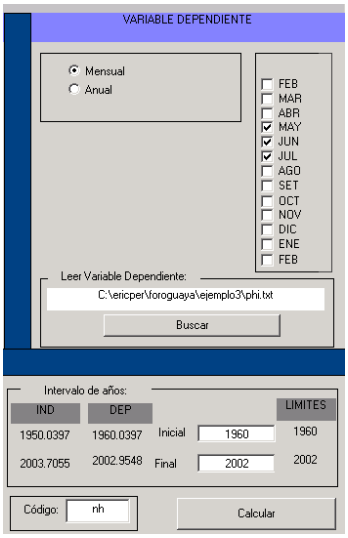
Para este tercer ejemplo usaremos los datos de precipitación de la estación Phillip Goldson (88.30° W, 17.53° N, 5 msnm), ubicada en el aeropuerto internacional de Belice, con registros mensuales de 1960 al 2002. Los datos se encuentran en el directorio ...\ejemplo3\phi.txt, y no contiene datos faltantes. El archivo phi.txt contiene dos columnas, la primera es el año mas la fracción del mes (vector de tiempo) y la segunda el acumulado de precipitación mensual en mm para la estación de Phillip Goldson:

```
%Año         Precip. Phillip Goldson
1960.039726   68.100000
1960.123056   20.300000
1960.206386   30.700000
1960.289717   95.000000
    ⋮            ⋮
2002.704804 106.400000
2002.788134 35.500000
2002.871464 97.800000
2002.954795 244.100000
```

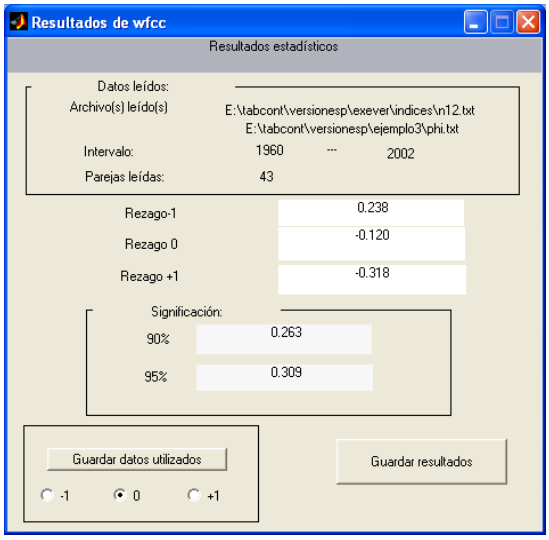
Exploraremos en este ejemplo la relación entre la precipitación de MJJ en Phillip Goldson y el Índice del Niño 1+2 para los periodos DEF, MAM, JJA, SON. Como primer paso al ejecutar el programa wfcc.exe, debemos marcar el círculo Externo Mensual bajo variable independiente, ya que el Niño 1+2 no es uno de los índices climáticos que se brindan con el programa, luego en los meses marcamos DIC, ENE y FEB, posteriormente hacemos clic en Buscar, para encontrar el archivo n12.txt que contiene los datos mensuales del índice Niño 1+2 y hacemos clic en abrir u open.



Seguidamente, bajo `variable dependiente`, marcamos `mensual`, ya que el archivo `phi.txt` contiene datos mensuales, luego en los meses marcamos `MAY, JUN` y `JUL`, posteriormente buscamos el archivo `phi.txt` en el directorio correspondiente y hacemos clic en `abrir` u `open`. Luego de esto elegimos el periodo a usar, para nuestro ejemplo 1960 en `Inicial` y 2002 en `Final`, también debemos digitar el `Código` del dato faltante, en este caso podemos poner `nh` ya que no tenemos datos faltantes.

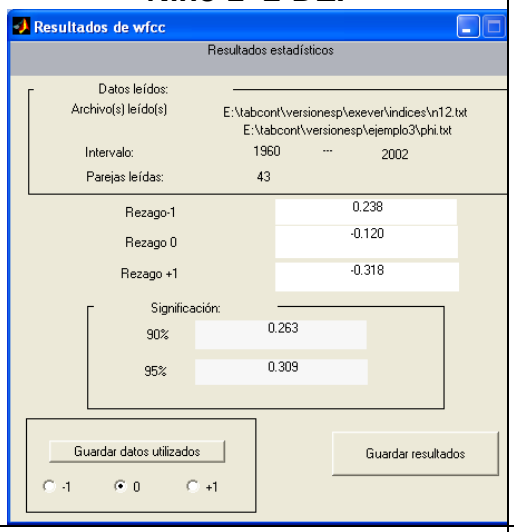


Una vez hecho lo anterior hacemos clic en `Calcular`, lo que despliega la siguiente pantalla:

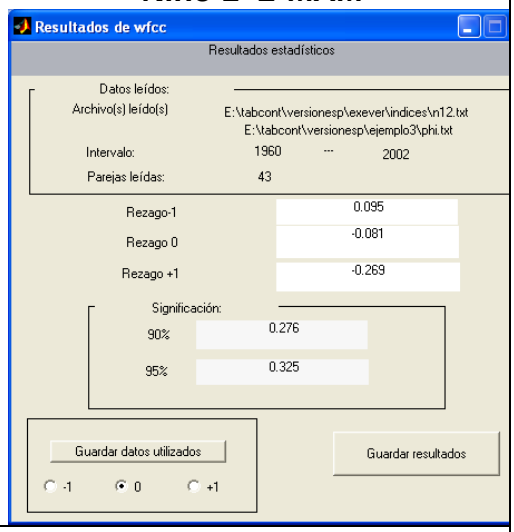


Estos resultados se refieren a la función de correlación cruzada o FCC entre Niño1+2-DEF del año anterior a la precipitación en Phillip Goldson durante MJJ o rezago –1, entre Niño1+2-DEF del mismo año a la precipitación en Phillip Goldson durante MJJ o rezago 0 y entre Niño1+2-DEF del año siguiente a la precipitación en Phillip Goldson durante MJJ o rezago +1. Donde Dic es el mes que define los años –1, 0 y +1. También se incluyen los niveles de significación de estos valores de estas correlaciones al 90 y 95%. Repitiendo los pasos anteriores pero seleccionando como periodos en la variable independiente MAM, JJA y SON obtenemos los siguientes resultados:

Niño 1+2-DEF

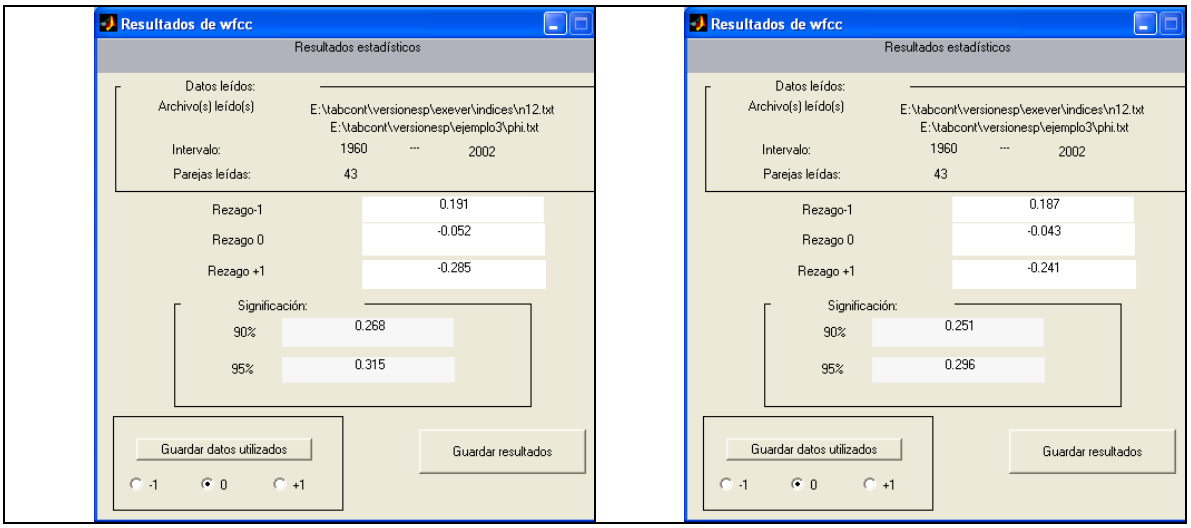


Niño 1+2-MAM



Niño 1&2-JJA

Niño 1&2-SON

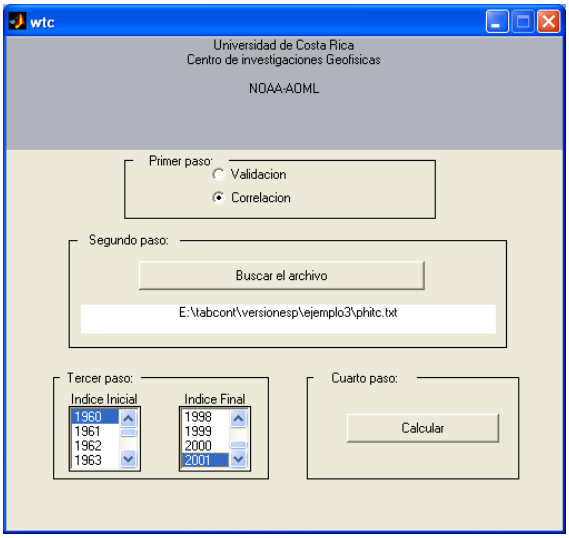


Nótese que el valor de la FCC entre el Niño1+2-DEF y la precipitación de MJJ del año 0 es 0.24, la cual no es significativa al 90%, sin embargo los otros valores altos de las correlaciones los encontramos para los rezagos +1, lo que nos indica que el evento de la precipitacion antecede, por varias estaciones climáticas, al del índice Niño1+2, esto es muy difícil de usar en un esquema predictivo; por otro lado recordemos que estamos usando los resultados de la FCC como una guia para el análisis de la tabla de contingencia y no para ajustar un modelo de regresión lineal. Como próximo paso entonces hacemos `clic` en el botón `Guardar datos utilizados` en la pantalla de los resultados del Niño1+2-DEF y los salvamos en el archivo `...\ejemplo3\phin12.txt`. El archivo `phin12.txt` contiene tres columnas, la primera es el año, la segunda el promedio de los valores del Niño 1&2 para el trimestre DEF y la tercera el promedio las anomalías de la precipitación en la estación de Phillip Goldon para el trimestre MJJ. Las líneas precedidas por el símbolo % se toman como comentarios:

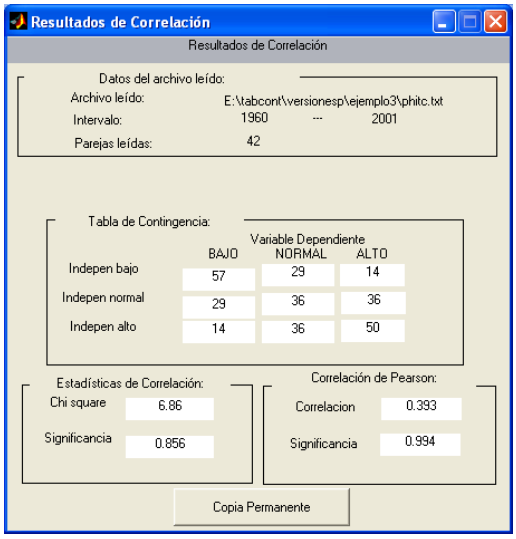
```
% Archivo de series anuales:
% Compuesto de los siguientes archivos fuente:
% Archivo leído
C:\ericper\viernes10oct\exever\indices\n12.txt
% Archivo leído C:\ericper\foroguaya\ejemplo2\phin12.txt
% Año inicial: 1960.00
% Año final: 2002.00
%        Año    Niño 1&2-DEF    Phillip Goldson, pcp, MJJ
  1.9600e+003   -1.2500e-001    1.6853e+002
  1.9610e+003    1.2333e-001    3.2227e+002
  1.9620e+003   -5.4667e-001    2.9987e+002
  1.9630e+003   -6.9000e-001    1.0563e+002
  1.9640e+003   -4.6333e-001    1.9810e+002
  1.9650e+003   -4.6333e-001    1.3870e+002
  1.9660e+003    4.2333e-001    3.0113e+002
  1.9670e+003   -5.4667e-001    1.4970e+002
```

```
1.9680e+003    -1.3700e+000     1.7667e+002
```

Luego de salvar este archivo ejecutamos el programa `wtc.exe`. En `Primer paso` selecionamos el círculo `Correlacion`, en `Segundo paso`, buscamos el archivo `phin12.txt`, en `Tercer paso` selecionamos 1960 en `Indice inicial` y 2002 en `Indice Final`:



Por último en `Cuarto paso` hacemos clic en `Calcular`, lo que despliega la siguiente pantalla:



Como parte del ejemplo, si observamos que se presento un valor por arriba de lo normal del Niño1+2 durante DEF ubicado en el tercer tercil, nuestra sugerencia para el foro climático, sobre las probabilidades esperadas en la precipitación de Phillip Goldson durante MJJ, será: 14% BN, 36% DN y 50% AN.

4.4 Aplicación a los foros climáticos IV

En este ejemplo tomaremos como variable dependiente un índice del promedio de los meses mayo, junio y julio (MJJ), de 1958 a 1998, de la temperatura superficial del aire en Centroamérica[10]. Este índice es representativo del comportamiento de este parámetro atmosférico en gran parte de la región como se muestra en la Fig. 4.4.1.

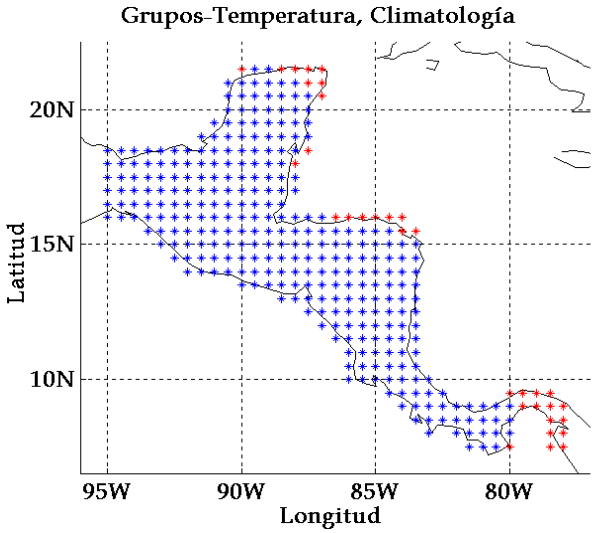


Fig. 4.4.1. Los asteriscos en azul representan los puntos en donde la curva de valores mensuales climatológicos de la temperatura superficial del aire presenta características similares al índice empleado en este ejemplo, en los puntos marcados con asteriscos en rojo no.[11] Esta EOF explica cerca del 80% de la varianza en la región.

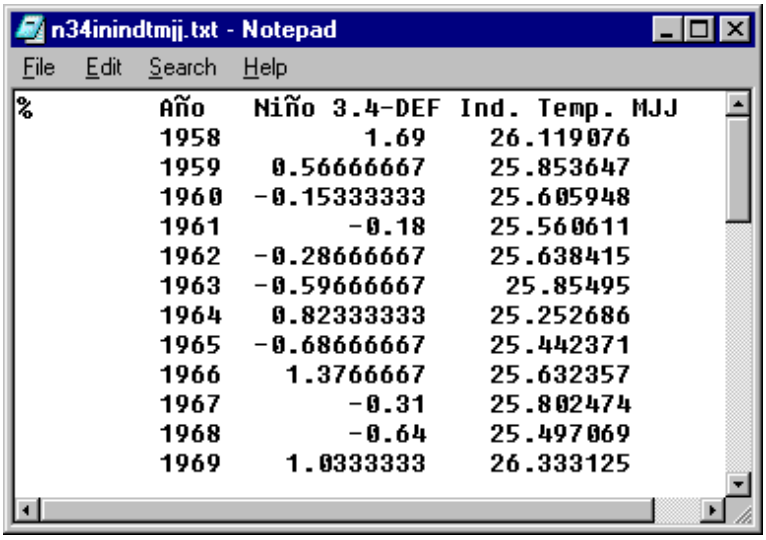
Los archivos a usar con el programa `wfcc.exe`, contenidos en el directorio `...\ejemplo4` son:

```
N34inindtmjj.txt
N34prindtmjj.txt
N34veindtmjj.txt
N34otindtmjj.txt,
```

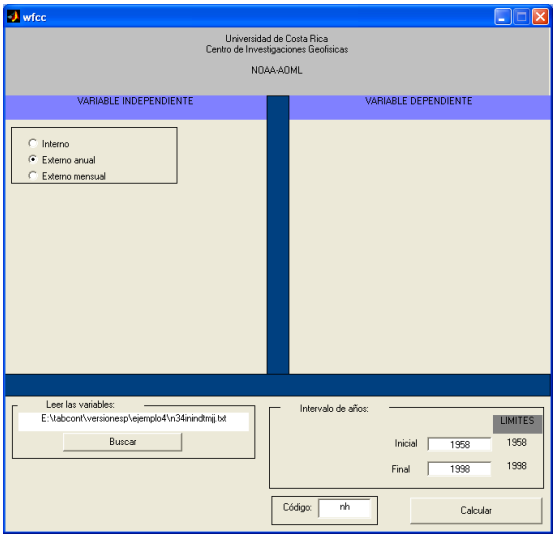
donde `N34` corresponde al índice de temperatura superficial del mar, para la región Niño 3.4; `in` es invierno (DEF), `pr` es primavera (MAM), `ve` es verano (JJA) y `ot` es otoño (SON); todas para el hemisferio norte. La primer columna de estos archivos es el año, la segunda es el correspondiente valor de Niño 3.4 y la tercera es el índice de temperatura para MJJ.

[10] El índice se construyó usando el análisis de componentes principales. Datos obtenidos como colaboración del proyecto CRN073-IAI.

[11] Ver: Alfaro, E., 2000: Response of Air Surface Temperatures over Central America to Oceanic Climate Variability Indices. *Tópicos Meteorológicos y Oceanográficos*, 7(2), 63-72.

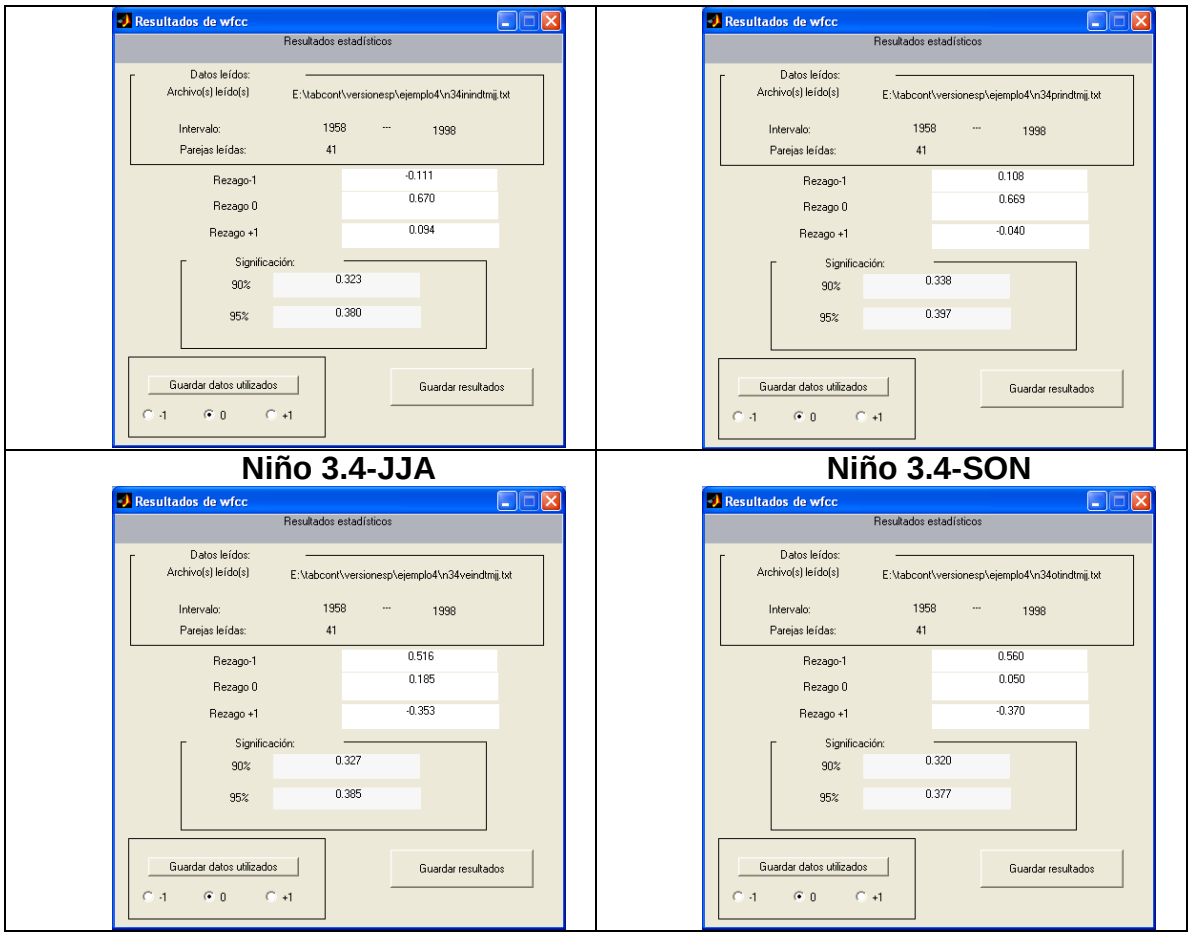


Debido a que estos archivos ya tienen ordenadas las tres columnas para estimar la FCC, como primer paso al ejecutar el programa `wfcc.exe`, debemos marcar el círculo `Externo Anual` bajo `variable independiente`, posteriormente hacemos `clic` en `Buscar`, para encontrar los archivos correspondientes y hacemos `clic` en `abrir` u `open`. Luego de esto elegimos el periodo a usar, para nuestro ejemplo 1958 en `Inicial` y 1998 en `Final`, también debemos digitar el `Código` del dato faltante, en este caso podemos poner `nh` ya que no tenemos datos faltantes.



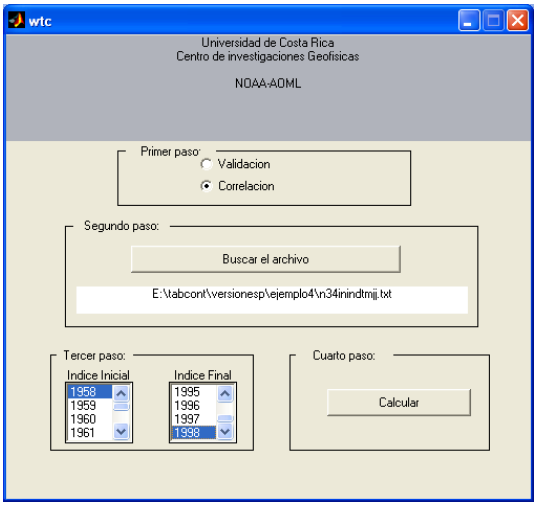
Una vez hecho lo anterior hacemos `clic` en `Calcular`, repitiendo el proceso para los cuatro archivos correspondientes a las cuatro estaciones climáticas del año del índice Niño 3.4, obtenemos lo siguiente:

Niño 3.4-DEF	Niño 3.4-MAM

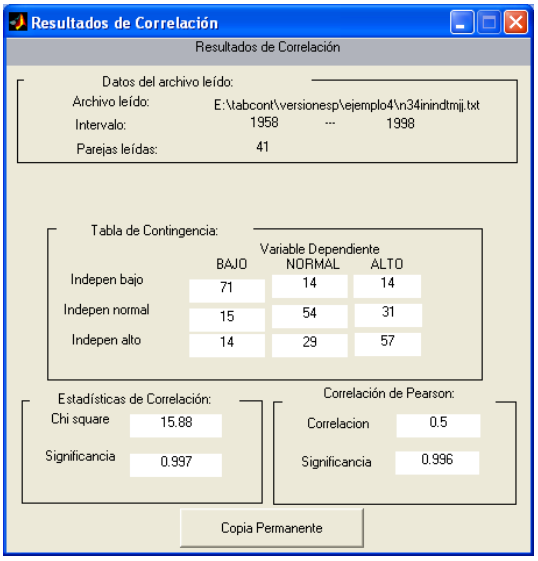


Estos resultados nos muestran que la mayor correlación es de 0.67 para el invierno en el índice Niño 3.4, durante el año 0 y nuestro índice de temperatura para MJJ. Además ese valor es significativo al 95%. En este ejemplo en particular el año 0 para DEF esta definido por Ene y Feb en lugar de Dic.

El siguiente paso sería ejecutar `wtc.exe`. En `Primer paso` selecionamos el círculo `Correlacion`, en `Segundo paso`, buscamos el archivo `N34ininditmjj.txt`, en `Tercer paso` selecionamos 1958 en `Indice inicial` y 1998 en `Indice Final`:



Por último en `Cuarto paso` hacemos `clic` en `Calcular`, lo que despliega la siguiente pantalla:



Esta relación sugiere un esquema de predicción clásico, ya que la variable independiente, Niño 3.4-DEF, antecede a la variable dependiente, Índice de temperatura en MJJ, por lo que para escoger el escenario en la variable independiente simplemente observamos en cual tercil se ubicó la variable independiente durante el último DEF. Si observamos que el Niño 3.4 durante el pasado DEF estuvo por arriba de lo normal o en el tercel tercil, nuestra sugerencia para la temperatura superficial del aire durante MJJ en el foro climático sería: 14% BN, 29% DN y 57% AN.

5. Información sobre los autores

Dr. F. Javier Soley (programación y principios estadísticos)
CIGEFI-Escuela de Física
Universidad de Costa Rica

2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320
Fax: (506) 234-2703
Email: fjsoley@racsa.co.cr

Dr. Eric J. Alfaro (principios estadísticos y aplicaciones prácticas en los RCOFs)
CIGEFI-Escuela de Física
Universidad de Costa Rica
2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320
Fax: (506) 234-2703
Email: ejalfaro@cariari.ucr.ac.cr, eafaro@ariel.efis.ucr.ac.cr

Dr. David B. Enfield (organización y aplicaciones prácticas en los RCOFs)
NOAA.AOML/PhOD
4301 Rickenbacker Causeway
Miami, Fl 33149
Email: David.Enfield@noaa.gov

La conversión de las funciones a C++ y la programación de las interfaces gráficas al usuario fue realizada por la estudiante de Ingeniería Eléctrica de la UCR:

Vilma Aguilar
CIGEFI
Universidad de Costa Rica
2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320
Fax: (506) 234-2703
Email: vilma@scratchy.emate.ucr.ac.cr

Nota de advertencia: Debido a que este software es distribuido libre de cargo, no se ofrece garantía de ningún tipo, ni explicita ni implícitamente. Los autores no se hacen responsables por el uso del mismo. Sin embargo están en la mejor disposición de contestar cualquier consulta que tenga el usuario sobre el material presentado, así como el considerar las sugerencias que se deseen hacer sobre el mismo, por lo que agradecemos su contacto con nosotros.

ADVERTENCIA GENERAL. Las diferentes versiones del software (ver Apéndice) provén diferentes estadísticos de diagnóstico que ayudan al usuario con la significancia de las relaciones predictor-predictante encontradas. En última instancia, la decisión de que tanta credibilidad se le puede otorgar a este resultado depende de su análisis. Aquellos usuarios con más experiencia en los principios estadísticos utilizados están

menos propensos a cometer errores de juicio o hacer proyecciones injustificadas. Se le recomienda al usuario estudiar detenidamente el manual y los ejemplos incluidos, especialmente en referencia a los estadísticos de diagnóstico. Tomando esto en consideración, se les invita a hacer un uso máximo de ellos.

6. Agradecimientos

Este trabajo se desarrolló gracias al apoyo de la NOAA-OGP a través del proyecto *A special proposal to improved regional climate Outlooks in Latin America* y de la Universidad de Costa Rica a través de los proyectos VI-112-99-305, ED-1040 y MM5-UCR-CRRH.

Apéndice.

Descripción de la evolución de los programas de la tabla de contingencia.

HISTORIA

1-) Abril -2003

Dos programas: corcruz3.exe y valdos.exe que corren desde la línea de comandos de DOS. Estos fueron presentados y distribuidos durante el COF de San Pedro Sula en Abril de 2003. No se le conocen errores de calculo o funcionamiento (pulgas) hasta el momento para esta versión. Este versión está disponible solo en español y para su uso en DOS (sin interfase gráfica).

Corcruz3.exe calcula tres rezagos (0, ±1) de la función de correlación cruzada y los errores estándar de rezago alto al 90 y 95 % de nivel de confianza. Los archivos de entrada en formato texto consisten de tres columnas con un índice temporal, variable independiente anual y variable dependiente anual.

Valdos.exe calcula la tabla de contingencia 3x3 entre la variable independiente y dependiente. Tiene dos opciones; validación y correlación. La primera ofrece una serie de estadísticos para cuantificar el grado de la relación entre dos variables con correlación positiva. La segunda opción se puede utilizar para variables con correlación positiva o negativa. La entrada de los datos es similar a la de corcruz3.exe.

Estadísticos calculados en la opción validación

a)Razón de aciertos
b)Puntaje de destreza
c)Falsas alarmas arriba y debajo de la normal
d)Porcentaje de detección arriba y debajo de la normal
a)Correlación de Pearson y su significancia
e)Prueba Chi cuadrada con su significancia
f)G cuadrada con su significancia
g)Error lineal en el espacio de probabilidad

Estadísticos calculados en la opción correlación

b)Correlación de Pearson y su significancia
c) Prueba Chi cuadrada con su significancia.
d)Pueba G cuadrada con su significancia.

Limitaciones: los usuarios, especialmente aquellos que solo han usado MS Windows, tienen dificultad al utilizarlos por el desconocimiento del ambiente DOS

2-) Noviembre -2003

Primera versión en ambiente de ventanas presentado en el COF de Guayaquil de Noviembre 2003. Los programas se llaman wfcc.exe y wtc.exe. Estos son los sucesores de corcruz3.exe y valdos.exe, respectivamente. Para este momento el software estaba disponible todavía solo en español.

Wtc.exe tiene las mismas funciones que valdos.exe, sólo que calcula sólo una de las pruebas chi cuadrada o g cuadrada y su respectiva significancia. Si no hay valores en la tabla cercanos a cero calcula la G cuadrada, y si los hay, la Chi cuadrada.

Tiene una pulga conocida: si se pulsa el botón de cancelar en la ventana para nombrar archivos, el programa aborta y se debe reiniciar el mismo.

A wfcc.exe se le añadieron varias funciones. Además de admitir una entrada de datos anuales como la de corcruz3.exe, permite la introducción de la variable independiente y dependiente como registros mensuales. De esta forma el usuario se ahorra un paso, como una opción, al no tener que calcular sus datos anuales de los registros mensuales. Incluye también algunos de los índices mensuales más utilizados en estudios climáticos como una opción por defecto, sin que los usuarios deban aportar estos datos. El usuario especifica que meses consecutivos desea que se promedien para construir las series anuales. Calcula los mismos estadísticos de diagnóstico que corcruz3.exe. Permite guardar en un archivo texto los datos usados en el cálculo de la correlación correspondiente a rezago cero.

Pulgas conocidas:

a)Si se pulsa el botón de cancelar en la ventana para nombrar archivos, el programa aborta.
b)La barra de avance en la ventana de espera no avanza. Provoca mensajes de advertencia.
c)No maneja correctamente la presencia de datos faltantes en las opciones de variable independiente y dependiente mensuales.
d)Cuando en las opciones de variable mensual se marcan meses que abarcan dos años, la correlación cruzada calculada tiene una sesgo extra.
e)Despliega como número de parejas leídas el número total de líneas del archivo de entrada.

3-) Enero -2004

Esta fue la primera versión en inglés y español y fue preparada para los angloparlantes que participaron para el COF de Kingston, Jamaica en marzo del 2004. Es esencialmente la misma versión de noviembre - 2003 con las siguientes pulgas corregidas del programa wfcc.exe (o wfcceng.exe en la versión en inglés):

a)Maneja correctamente la presencia de datos faltantes en las opciones de variable independiente y dependiente mensuales.
b)Si en las opciones de variable mensual se marcan meses que abarcan dos años, la correlación cruzada se calcula correctamente.
c)Despliega correctamente el número de parejas leídas.
d)La barra de avance ya no provoca mensajes de advertencia.

Persisten las siguientes pulgas:

f)Si se pulsa el botón de cancelar en la ventana para nombrar archivos, el programa aborta.
g)La barra de avance en la ventana de espera no avanza.

Se añadió la opción de guardar los datos utilizados en el calculo de la correlación con rezago -1,0 o 1. El archivo creado se puede utilizar en wtc.exe (o wtceng.exe en la versión en inglés).

Wtceng.exe no sufrió cambios.

4-) Diciembre, 2004 - Enero, 2005.

Wfcc.exe no sufrió cambios.

En wtc.exe se corrigió el cálculo y presentación en pantalla de chi2 y su significancia. También se detectó y corrigió un error menor de redondeo en el cálculo de las probabilidades de los escenarios neutrales, sobre todo cuando se usaban muy pocos datos o hay muchos datos repetidos, pe. época seca. Nótese que en estos casos el análisis contingente no es recomendable del todo. Se añadió en la pantalla de salida del DOS la tabla de contingencia en frecuencias absolutas, lo cual es útil al usuario. Cuando el número de pares de datos usados es menor a 45 el programa despliega un mensaje de advertencia en la pantalla del DOS. Se incluyó en esta misma pantalla el despliegue de la tabla de contingencia de las frecuencias absolutas que algunos usuarios han sugerido como útiles. La significancia del estadístico de correlación de Pearson se calcula ahora en forma mas conservadora utilizando el error estándar de rezago alto que toma en cuenta la autocorrelación de las series. Se obtiene así una mejor idea de la significancia de las series, siendo ahora en general menor al cálculo anterior.

Use of a Contingency Table for Climate Applications

Eric J. Alfaro and F. Javier Soley

School of Physics and Center of Geophysical Research

University of Costa Rica

David B. Enfield

Physical Oceanography Division

Atlantic Oceanographic and Meteorological Laboratory

National Oceanic and Atmospheric Administration

Index

5. Introduction

Since 1997, Regional Climate Outlook Forums (RCOFs) have taken place in various Latin American countries. They have been funded by several international agencies and with the assistance of local and regional entities such as the Regional Committee of Hydraulic Resources (CRRH) in Central America.

Generally, these forums gather representatives of the Meteorological and Hydrological services, as well as members of the scientific and academic community, who work on the elaboration of regional and local climate forecasts for the next 1-2 seasons. The objective of these forums is to use national climatic experience to elaborate a regional consensus for the climate outlook. Precipitation is the variable generally forecast for the months following the forum. The forecast is presented in a format that is useful for the agencies involved. The recommended methodology for the forecast is quite simple. The probabilities of precipitation or temperature terciles, i.e. the predictands, are extracted from a contingency table of two variables in which a climate index is used as predictor, such as the Southern Oscillation index (SOI) or an average of sea surface temperature (SST) (e.g., Niño 3, TNA). This forecast is later integrated geographically with the coordinated inputs from the countries of the region, and is used as a tool for the meteorological services and as a basis for expected impact scenarios for stakeholders and decision-makers.

The scientific and academic communities have discussed certain problems that arise during the development of the forums and how the research results can be better used to improve the forums' products. An application project funded by the Office of Global Programs of the National Oceanographic and Atmospheric Agency (OGP-NOAA) has been executed by the University of Costa Rica and the NOAA Atlantic Oceanographic and Meteorological Laboratory (hereafter the NOAA-UCR-PROJECT). This is an extension project that aims at fulfilling some of the national meteorological services' needs regarding climate forecasts, so their participation in the RCOFs can be more efficient and better coordinated between countries. One of the problems identified during the RCOFs is that because there is not an standardized methodology for producing the forecast, the contributions from different countries can result in a disjointed regional forecast that is some times physically inconsistent across political borders. Moreover, it appears that the statistics of the contingency table approach are not familiar to some of the participants, such that the national climate forecasts are sometimes based only on subjective evaluations. Some of the roots of these problems have been identified: i) the resources of some institutions are limited to the routine tasks and only a small portion of their budget is allocated for research and capacity building; and ii) there have been very few opportunities for training on the concepts required for the RCOFs.

Hence, the NOAA-UCR-PROJECT has created a series of user-friendly programs to facilitate the application of terciles in forecasting. These programs implement graphical user interfaces (GUIs). Early versions of the software have been distributed to the meteorological service personnel on a CD-ROM, which accompanies this manual, and which can be used in training sessions during the RCOF. It includes a library with a

series of specialized functions compiled under Matlab environment. Its primary target is the quantitative and categorical analysis of two variables using contingency table theory. This is accomplished by building a contingency table to determine the degree of association between the data sets as well as their conditional probability relationship. The Matlab functions were then compiled for the Windows environment, independent of the Matlab platform[12]. This manual is for the latest software version, and the history of the software versions is found in the Appendix.

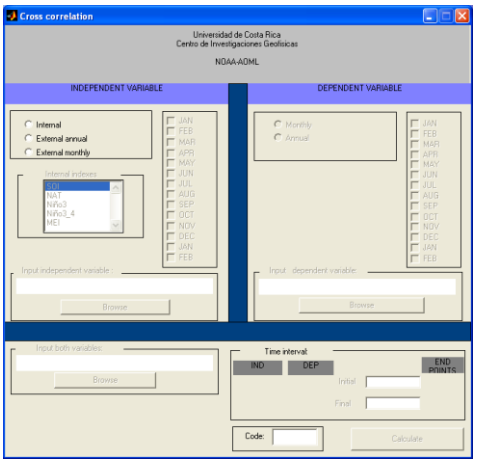
The main program (wtc.exe or wtceng.exe, Spanish and English software versions, respectively) can be used in two ways, the first to make an exploratory analysis between the two variables that are suspected to be correlated in time, positively or negatively (e.g., the precipitation at San Jose versus the SOI) and the second to validate a resulting forecast, that is, to examine the contrast between the observed and forecast patterns. The other program is wfcc.exe or wfcceng.exe (again, Spanish and English software versions, respectively), which makes the biased cross-correlation between two variables and allows to choose an appropriate predictor.

6. Program execution

The programs are executable under Windows environment. First, the user must copy the exever directory from the CD to the hard drive. In this directory, the user can find the executable programs wfcc.exe and wtc.exe (or wfcceng.exe and wtceng.exe). Afterwards, the user will create a shortcut for these programs on the computer desktop, which will avoid the accidental erasing of the libraries included in the exever directory, which are necessary for program execution.

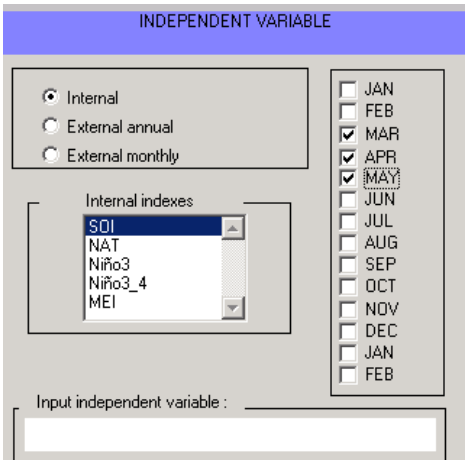
We will discuss first the wfcc.exe or wfcceng.exe, which calculates the biased cross-correlation between two variables for -1, 0, and +1 lags. This program is primarily of help for predictor identification. The two variables on which the analysis is based must be annual series, although the series themselves can be either annual or monthly time series. Here, the negative lag means that the independent variable leads the dependent variable. When executing the program the following graphical interface appears:

[12] An early version that works in MS-DOS environment is also included in the CD, as well as the respective User's Manual in Spanish, see the Appendix for more details.



The variables to be analyzed must be in separate ASCII or simple text files, in which the values of the columns of each row must be separated by one or more blank spaces. The first column is a sequential index and the second, the independent or dependent variable[13]. Any lines such as header lines that are considered comments must begin with a % sign, and the missing data must be codified appropriately with a numeric value, e.g., -9999.

This graphical interface allows the user to import the data in several ways. The independent variable (predictor variable) can be read in three different ways. If the user selects Internal, it means that he/she is going to use one of the climate indices that are already incorporated in the program, in which case, no user-supplied file is required. The user must select one of them. The five text archives (*.txt) corresponding to these indices, SOI, TNA, Niño3, Niño 3.4 and MEI, are in the directory c:\exever\indices. Afterwards, the user must select the season of the year that is going to be analyzed for this index, e.g. select SOI and then MAM (March, April, May).[14]

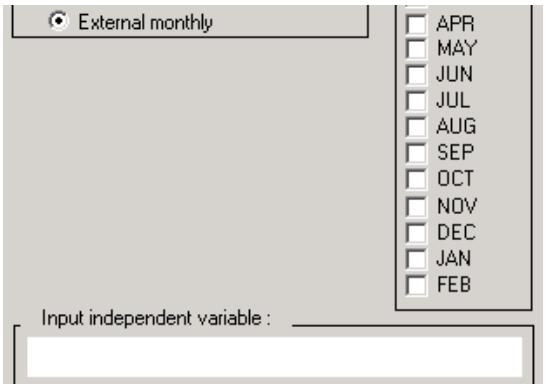


[13] An excel macro that transform the usual format of monthly data (eg. A matrix with N years and 13 columns: year, Jan, Feb, …, Dec) to a monthly data vector needed by wfcc.exe and wfcceng.exe is included in the file mattira.xls.

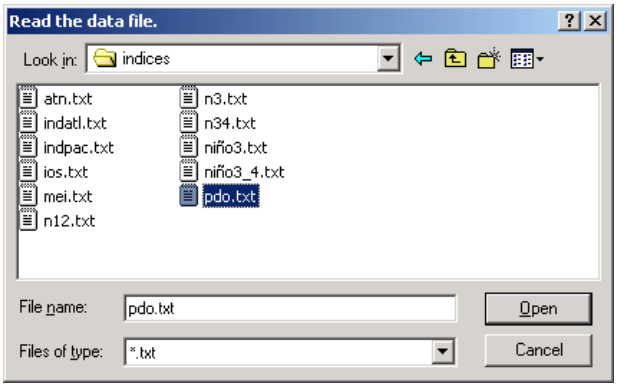
[14] Ascii files for Niño1+2, PDO and AMO indices are also provided in that directory. They could be used by the program trough the External Monthly option, described later in this section.

Next, the months corresponding to the season to be used for the predictor are checked from the list on the right. Then, the user will select Load, to move the annualized season predictor data into memory.

If the user selects External Monthly, it means that a climate index not included in the menu will be used and that it is in a text file with two columns of monthly values. The first column corresponds to the time axis and the second one to the climatic index. The file search must be made by clicking on the Browse button,



after which the user must select the file to use, clicking on OPEN,



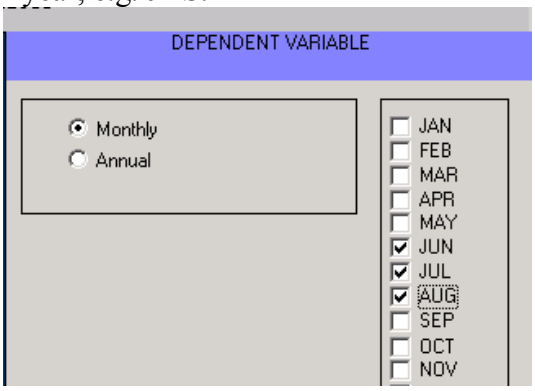
As in the previous case, the user must select the months corresponding to the season of the year to analyze for the independent variable, and the program creates the annualized sequence from the monthly data. Notice that if the user puts the directory exever with a different path than c:\exever, the indices displayed using the Internal option will not be found by the program but they can still be used trough the External monthly option.

When the user selects External Annual in the menu of the Independent Variable, the variables that will be analyzed are already annualized and must be contained in an ASCII file or simple text file with three columns. The first column is a sequential index (in most of the cases it is corresponding to the year, but not necessarily), the second column contains the independent variable and the third column is the dependent variable. It should be emphasized that in all the cases the lines that are considered comments must

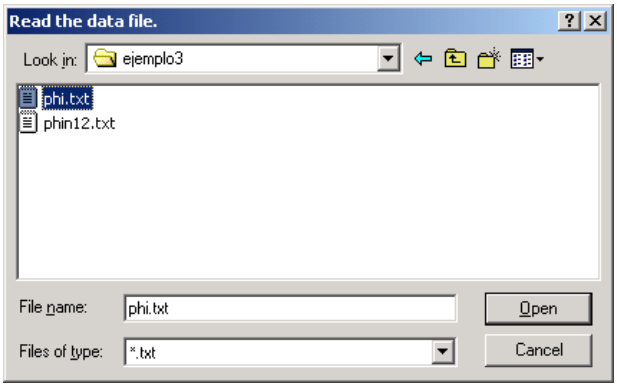
begin with a sign of % and as was mentioned previously, the missing data must be codified appropriately with a numerical value. For example[15]:

```
% Year            SOI, JJA(-1)      Quepos, IELL
1.9410000e+003  -1.7733333e+001  1.9000000e+001
1.9420000e+003  -1.8033333e+001  2.1000000e+001
1.9430000e+003  3.7666667e+000   2.7000000e+001
1.9440000e+003  9.3333333e-001   2.2000000e+001
1.9450000e+003  -3.1666667e+000  2.3000000e+001
1.9460000e+003  7.8333333e+000   2.4000000e+001
1.9470000e+003  -8.0666667e+000  -9999
1.9480000e+003  6.4000000e+000   -9999
1.9490000e+003  -2.7333333e+000  2.4000000e+001
```

The loading of the dependent (predictand) variable can be done in two ways. The first step is to select Monthly or Annual, depending on the nature of the data. If Monthly is selected, again it will activate the selection of the months that allow choosing a season or series of months for each year, e.g. JAS.



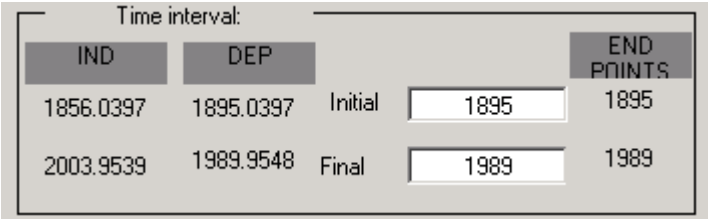
As in the previous case, the Browse button will allow the user to select the file for the analysis,



The Annual option of the Dependent Variable allows the selection of time series of annualized data, in which the text file contains two columns: the first column with the sequential index (years usually) and second with the data of the parameter to analyze.

[15] The format of the input data are not necessary exponential like in this example, see section 4.

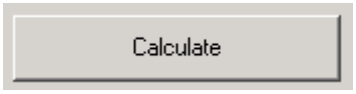
Once the user has the two variables in memory, independent and dependent, the interval of the sequential index with which he/she wishes to work must be selected. For this, we display the maximum and minimum values of the variables used, on the left hand side of the graphical interface. The period that the user selects to work with must be a subgroup of this common interval.



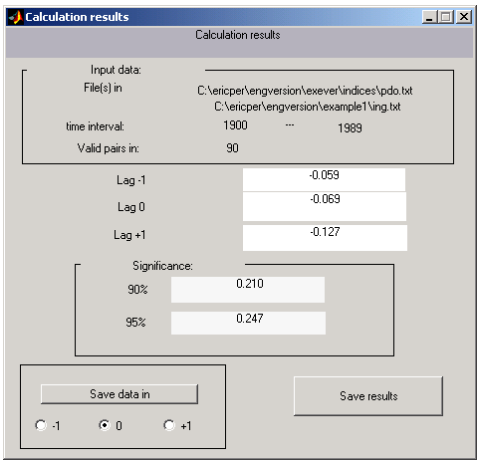
This analysis allows the existence of a reasonable number of missing data. If there are any missing data, the user must enter the proper Code that corresponds to missing data in the files use for this analysis,

,

or put nh if there aren't missing data. Next, to run the analysis the user must press Calculate,



This action will display a screen with the analysis of the cross-correlation function. The interpretation and use of these results will be explained in a later section.



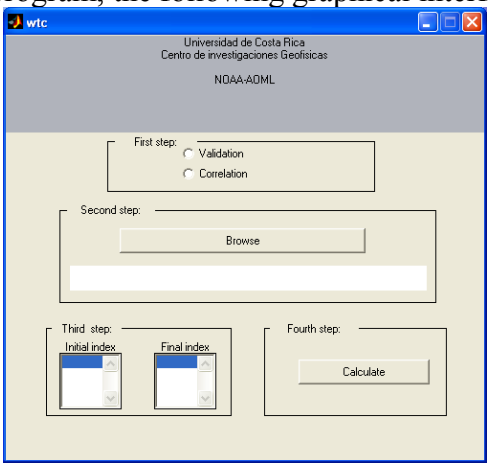
Finally, if the user wishes to save the results of the analysis in a text file, press Save results, also, if the user wishes to save the data for the contingency analysis in a text file, Save data in should be pressed, using the appropriate lag for the predictor that will be used for the contingency analysis, e.g. –1 for Lag –1, the year before, 0 for Lag 0, the same year or +1 for Lag +1 or the year after.

The second of the programs is wtc.exe or wtceng.exe. It builds the contingency table between two variables. The variables that will be analyzed must be contained in an ASCII

file in which the first column is a sequential index, the second column is considered the independent variable and the third column is the dependent variable. If there are one or two or more consecutive missing data, these must be eliminated from the analysis by preceding them with the comment symbol (%). For example:

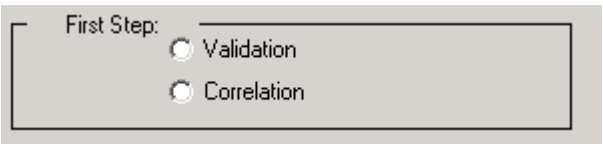
```
%      Year            SOI, JJA(-1)      Quepos, IELL
   1.9410000e+003  -1.7733333e+001   1.9000000e+001
   1.9420000e+003  -1.8033333e+001   2.1000000e+001
   1.9430000e+003   3.7666667e+000   2.7000000e+001
   1.9440000e+003   9.3333333e-001   2.2000000e+001
   1.9450000e+003  -3.1666667e+000   2.3000000e+001
   1.9460000e+003   7.8333333e+000   2.4000000e+001
%  1.9470000e+003  -8.0666667e+000   -9999
%  1.9480000e+003   6.4000000e+000   -9999
   1.9490000e+003  -2.7333333e+000   2.4000000e+001
```

When executing this program, the following graphical interface is displayed:

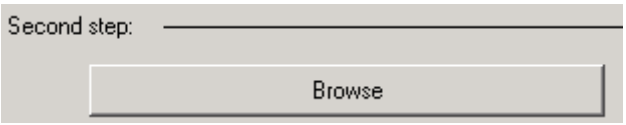


Once this interface is displayed, the following steps are required:

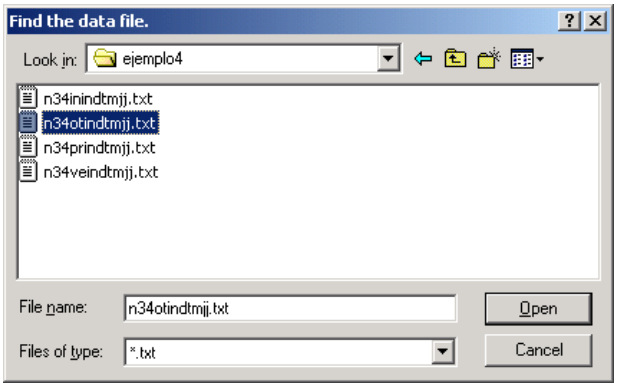
6) Select Validation: to do a cross-validation in order to contrast the behavior of data that results from a forecast with observed data; or select Correlation to make the exploratory analysis between two variables that we assume to be *negatively or positively* correlated in time (e.g. Precipitation in San José in MJJ versus Southern Oscillation Index in DJF).



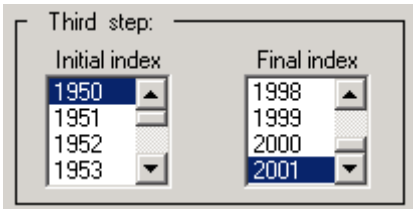
7) Select the file to be analyzed by pressing Browse,



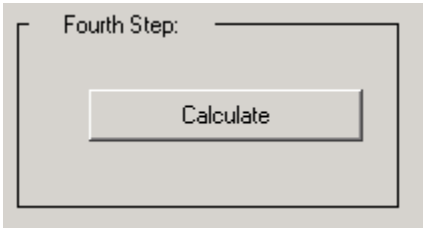
Once selected, the user must press open to read it to the memory.



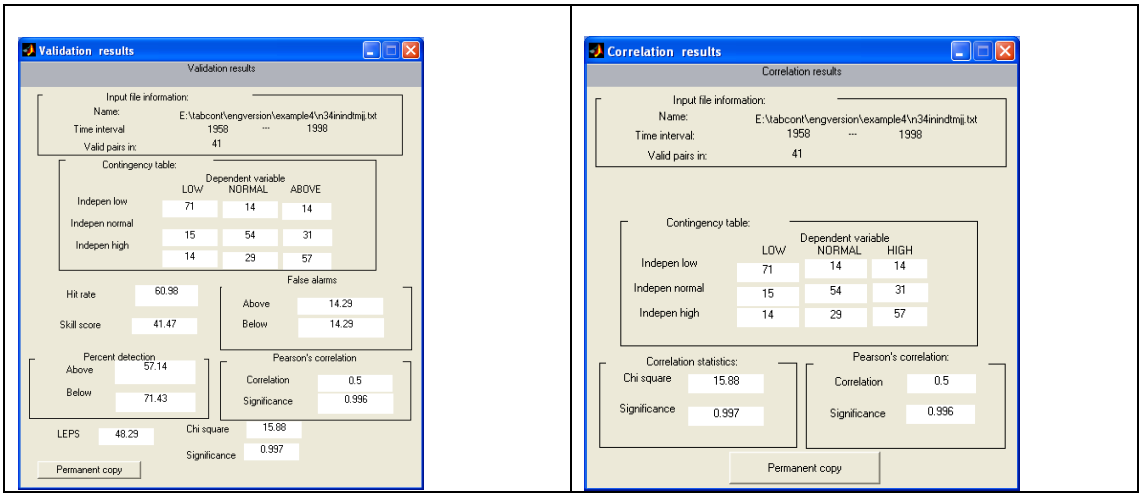
8) Choose the analysis period clicking on Initial index and Final Index; the last one must be greater than the first one.



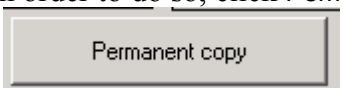
9) The contingency analysis can now be done by pressing Calculate.



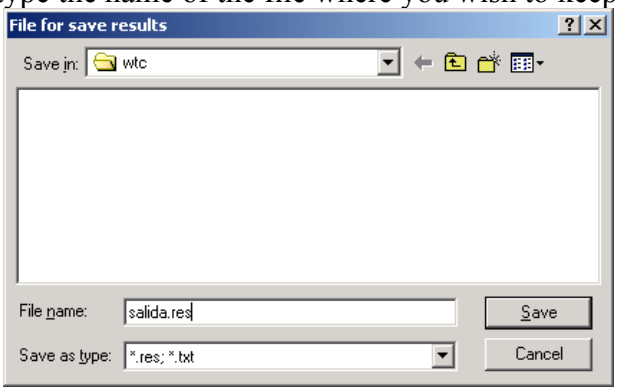
The program will display the contingency table and various statistical results from the analysis, which will be explained later in section 3. The table format will depend on the option that the user chosed in the first step, Validation or Correlation.



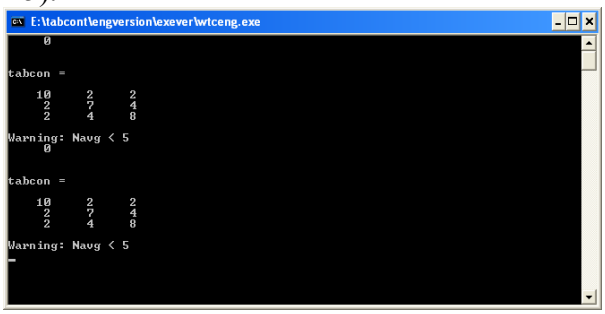
10) This step is optional. In this step the results from the analysis will be saved in a text or ASCII file. In order to do so, click Permanent copy,



and then type the name of the file where you wish to keep the results.



Notice that you can choose the folder where you'll keep the outcome file, too. Contingency Table results for the absolute frequencies are displayed on the MS-DOS screen, and a warning is given if the total number of pairs are less than 45 (average cell pairs is less than 5).



7. Theoretical aspects of the program and interpretation of results

3.1 Cross-correlation function

The key for the application of the contingency table is to choose dependent and independent variables that have a significant predictive relationship. In other words, we want them to be well correlated. A mathematical tool that allows us to quantify the common behavior (or the degree of common information) between two variables is the cross-correlation function, which is defined as follows:

$$c_{xy}[k] = \frac{1}{N}\sum_{t=1}^{N-k}\frac{\left(x[t]-\bar{x}\right)\left(y[t+k]-\bar{y}\right)}{\sigma_x\sigma_y} \qquad k = 0,1,2\ldots,$$

$$c_{yx}[k] = \frac{1}{N}\sum_{t=1}^{N-k}\frac{\left(y[t]-\bar{y}\right)\left(x[t+k]-\bar{x}\right)}{\sigma_x\sigma_y} \qquad k = 0,1,2\ldots,$$

Where $\bar{x}$ and $\bar{y}$ are the sampled means of the series *x* and *y*. Denoting the demeaned residuals of the series *x'* and *y'* we can simplify the equation to

$$c_{xy}[k] = \frac{1}{N}\sum_{t=1}^{N-k} x'[t]y'[t+k] \qquad k = 0,1,2\ldots,$$

$$c_{yx}[k] = \frac{1}{N}\sum_{t=1}^{N-k} y'[t]x'[t+k] \qquad k = 0,1,2\ldots,$$

From these equations consider the following:

h) We will call the product *x'[t]y'[t+k]* the product of *x'* with *y'* lagged by *k* sampling intervals. Then, $c_{xy}[k]$ is the mean of the products between *x'* and *y'* lagging *k* sampling intervals. In the same way $c_{yx}[k]$ is the mean of the products between *y'* and *x'* lagging *k* sampling intervals. For example: lets assumed that *x'* are the residuals of the Southern Oscillationn Index (SOI) and *y'* the residuals of a monthly precipitation (PREC) time series. $c_{xy}[0]$ is the average of the products SOI[Jan] PREC[Jan], SOI[Feb] PREC[Feb], SOI[Mar] PREC[Mar], etc. $c_{xy}[1]$ is the average value of the products SOI[Jan] PREC[Feb], SOI[Feb] PREC[Mar], SOI[Mar] PREC[Apr], etc. $c_{xy}[2]$ is the average value of the products SOI[Jan] PREC[Mar], SOI[Feb] PREC[Apr],SOI[Mar] PREC[May], etc.

i) Let's assume that *x'[t]* and *y'[t+k]* do not have a common behavior, meaning that when *x'[t]* is positive, *y'[t+k]* can be positive or negative with equal probability and when *x'[t]* is negative, *y'[t+k]* can also be positive or negative. Then the mean value tends to zero and we say that both series are not correlated at lag *k*.

j) Suppose that *x'[t]* and *y'[t+k]* have a common proportional behavior. This means that when *x'[t]* is positive, *y'[t+k]* is also preferably positive and when *x'[t]* is negative, *y'[t+k]* is preferably negative. In this case the average value of the product is positive and we say that the series are positively correlated at lag *k*.

k) Suppose that *x'[t]* and *y'[t+k]* have a common inverse behavior such that when *x'[t]* obtains a positive value, *y'[t+k]* preferably obtains negative values. And when *x'[t]* obtains a negative value, *y'[t+k]* preferably obtains positive values.

Then the average value is negative and we say that the series are negatively correlated at lag *k* or that they are anti-correlated at lag *k*.

l) The definition of $c_{yx}[k]$ can be extended to negative lag values. In our example $c_{yx}[-1]$ is the average value of the products SOI[Dec] PREC[Jan], SOI[Jan] PREC[Feb], SOI[Feb] PREC[Mar], etc. But that mean value coincides with $c_{yx}[1]$. In general, $c_{yx}[-k] = c_{xy}[k]$.

m) The value at *0* lag is associated with the covariance between the variables *x* and *y*.

n) When $x = y$, the cross-covariance coincides with the autocovariance function.

3.1.1 Significance Test

In the case that *x* and *y* are randomly uncorrelated series, the variance of the estimated cross-correlation function can be computed by

$$\sigma_{c_{xy}}{}^2 = \frac{1}{N}\sum_{m=-P}^{+P}\rho_X[m]\rho_Y[m].$$

ρ_X and ρ_Y are the autocorrelation functions of *x* and *y*, *N* is the nominal number of degrees of freedom, and σ is known as the large lag standard error. The significance levels are then 1.645σ, 2.0σ, and 2.58σ with 90%, 95% and 99 % confidence levels, respectively. If the estimated values of the cross-correlation exceed the significance levels, then the null hypothesis (i.e., that the series are not correlated) is rejected at the corresponding confidence level and then the alternative hypothesis is accepted (the series are correlated). The program wfcc.exe calculates the 90% and 95% confidence levels.

3.1.2 Practical Considerations

The experience gained from the Regional Climate Outlook Forums (RCOFs) is that correlations between monthly time series do not yield satisfactory results because the seasonal patterns are mixed up (for example, dry season with the rainy one, beginning of the rainy season with its end, etc.). This tends to lower the correlations and to make the data interpretation more difficult. For the purposes of the RCOFs it is more productive to correlate the average values of several consecutive months (typically three) of each year, changing the scale from monthly to annual such that only the same seasons within different years are considered. For example, the average Tropical North Atlantic (TNA) sea surface temperature (SST) can be correlated for the months of June, July and August with a precipitation index derived from a coherent grouping of Central America stations for the months of December, January and February (EOF-PCP). Figure 3.1.2.1 shows the cross-correlation function between these time series for 1958-1999 with –5 to 5 lags. The red and green lines define the significance levels at 90% and 95%, respectively. Note that there are two significant values at 95%: a positive value in the lag -1 (TNA leading EOF-PCP by one year), and a negative value at lag 0 (the two time series are in phase). The rest of the values are not significant at the 90% confidence level.

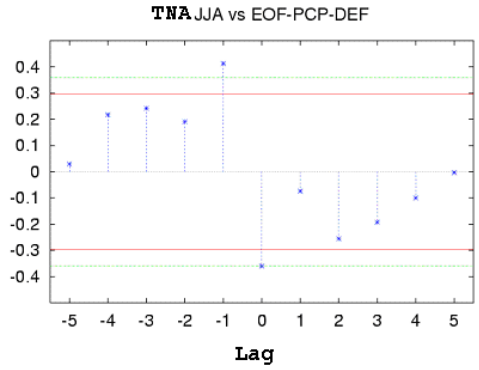


Fig. 3.1.2.1 Cross-correlation function between the mean TNA SST for the months of June, July and August with a precipitation index derived from several stations in Central America for the months of December, January and February (EOF-PCP)[16].

For the purposes of the RCOFs, there is no point in looking for correlations at lags greater than a year, therefore the program wfcc.exe only calculates the cross-correlation function for lags k = 0, ± 1.

3.2 Construction and Interpretation of the Contingency Table

The association between two continuous variables is determined by the distribution of their conditional probabilities (like the conditional probability for rainfall given that the Niño 3 index is in its upper tercile). In general, these distributions are unknown, and are substituted by a 'contingency table' derived from samples taken from the parent variable populations. In the case of the RCOF application, the samples are the monthly averaged time series of rainfall measurements at meteorological stations, and the climate indices derived from predictor variables (NIÑO3, MEI, TNA, etc.). To construct this table the continuous variables are divided into discrete categories, i.e. the first variable or independent variable X with M categories and the second variable or the dependent variable Y, with N categories. Each pair of values (x_i, y_i) belongs to one and only one of the M x N joint categories (i-th and j-th). Then the empirical frequencies f_{ij} are calculated. These are the number of pairs that belong to the category ij. If the association between the two variables is too weak, the frequencies of the M x N categories are similar, such that none are significantly more probable than others and there is little basis for prediction. In a 3-D plot the surface appears flat because the values of X within the category i can be associated with values that belong to any of the N categories of Y (Fig. 3.2.1.a). If the linear association is strong, the surface can get large values along one of the diagonals and low values on the corners that do not belong to the diagonal (Fig. 3.2.1.b). In the case of a positive association, high values appear along the main diagonal and in the case of a negative association, along the secondary diagonal.

[16] See: Alfaro, E., 2002: Some Characteristics of the Annual Precipitation Cycle in Central America and their Relationships with its Surrounding Tropical Oceans. *Tópicos Meteorológicos y Oceanográficos*, 9(2), 88-103.

a) b)

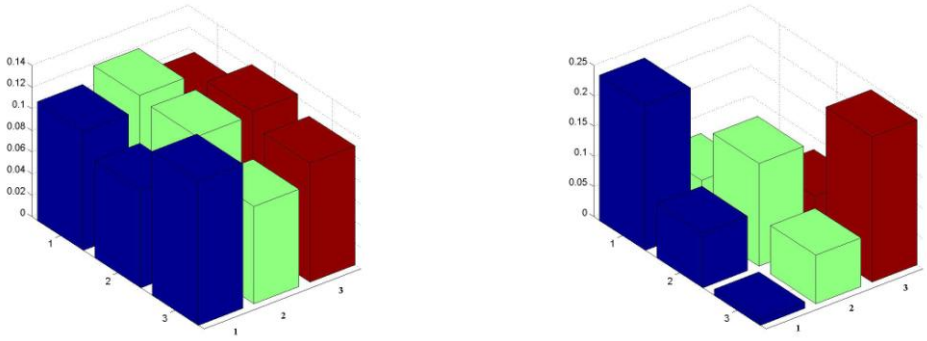


Figure 3.2.1. Empirical Probabilities where: a) the association between variables is weak. Note that the frequency in each one of the boxes is very similar. b) The association between variables is strong. Note that the frequencies along one diagonal are larger than the rest of categories.

By dividing the empirical frequencies by the total number of pairs we get the empirical probabilities. This means that $p_{ij} = f_{ij} / n$ is the empirical probability that a pair of values belongs to the joint category ij. At the Regional Climate Forums (RCOFs) the individual values of p_{ij} or f_{ij} can be expressed as percentages, P_{ij}, in relation to the M locations of the independent variable or predictor, i.e.,

$$P_{ij} = \frac{p_{ij}}{\sum_{j=1}^{N} p_{ij}} \times 100 \text{ , or, } \quad P_{ij} = \frac{f_{ij}}{\sum_{j=1}^{N} f_{ij}} \times 100 \text{ .}$$

The statistics used to quantify the degree of association in an objective way will be discussed again in section 3.4.

3.3 The *3* x *3 Contingency Table*

At first it would seem convenient to use a large number of categories to accomplish higher resolution. In practice, we have realized that a high number of categories is difficult to interpret because of the high number of probabilities to consider. Moreover, a large number of data would be needed to accomplish a stable analysis of many M x N categories. When dividing the two variables into 'terciles', this means, doing $M = N = 3$, or that *9* joint categories are obtained, allowing a certain degree of resolution and a manageable number of possibilities. On the other hand, a more pragmatic way of focusing the selection of *3* x *3* is that frequently a prediction can be more easily

understood by the public in general using concepts like: *normal, above normal and below normal.*

Let's assume that we are trying to do a cross-validation between an observed climate index (independent variable, or predictor) and a group of values for which we wish to project a future outlook (dependent variable, predictand). As an example we could have as a predictive variable some index related to El Niño/Southern Oscillation – ENSO- (i.e., SOI). Meanwhile the variable to predict could be the precipitation we expect in some climatic region or country basin, normally measured by the average of a group of meteorological stations. The observations are then divided into terciles: low values (Obs. Low), Normal values (Obs. Normal) and high values (Obs. High). Each observation (or row of the data file) has to belong to only one category (the categories are exclusive) and the three categories comprise all of the possibilities in the "sample" or group of observations (the categories are exhaustive). This is illustrated in the Venn diagram of the Figure 3.3.1. The Figure 3.3.2 shows the Venn diagram of the predictions. Since the categories are terciles, each category is a third of the total area. When considering the pairs of the Venn diagram for the pairs (observed values, predicted value), it takes the form of Figure 3.3.3 in which *9* excluding and exhaustive categories are identified for the pairs. For example, a low value of the observation can be associated with a low value (OB-PB), a normal (OB-PN) or a high one (OB-PA) of the predicted value.

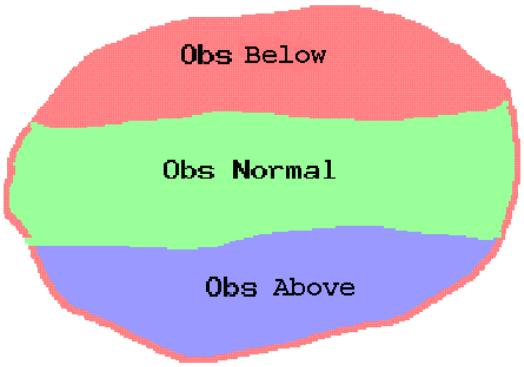


Figure 3.3.1. Venn diagram for the independent variable categorized in terciles.

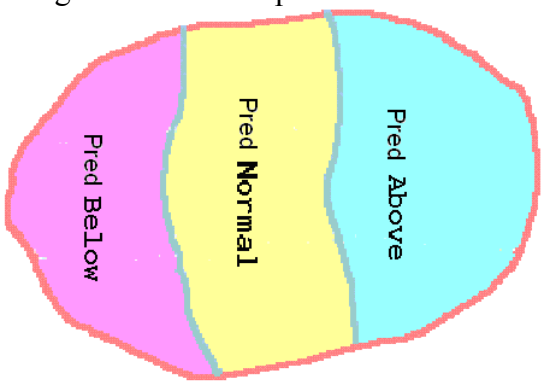


Figure 3.3.2. Venn diagram of the dependent variable categorized in terciles.

In other words, for each of the two variables we have determined the historical categories to be considered normal, above normal and below normal, in the absence of any other criteria. But, if we suspect that the category of the predictor variable affects the

distribution of precipitation (e.g., more rain with La Niña, less rain with El Niño), we can test this and accept or reject it by distributing the data amongst the joint categories.

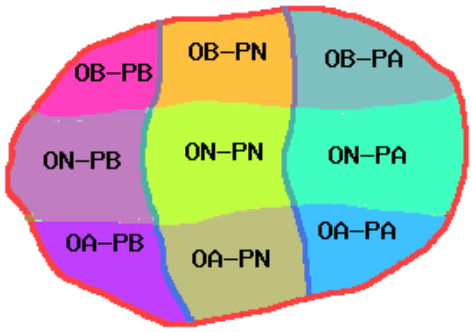


Figure 3.3.3. Venn Diagram for the joint categories.

Figure 3.3.3 shows that the area of each one of the *9* possibilities is approximately of the same value, based on their respective areas. This happens when the association between observations and predictions is weak or random (similar to the shown in Fig. 3.2.1.a). In this case the condition of the predictor variable has no significant effect over the variable to be predicted. In other words, and referring to the previous example, when the Pacific Ocean is in a normal state, or Niño or Niña, the distribution of the rainfall probabilities do not depart in a significant way from the overall climatological expectations (Figure. 3.3.2): [1/3, 1/3, 1/3]. This hypothetical case of statistical independence between variables is used as a point of reference to evaluate the goodness of a prediction; we refer to it as the 'null hypothesis'. The more the statistical properties of a contingency table differ from the properties of the contingency table for the null hypothesis (Figure 3.3.3), the stronger is the association between the independent and dependent variable. Many of the statistics and the significance tests in these programs are used to quantify how different a given contingency table is from the contingency table of independent variables.

For independent events, the probability that they occur simultaneously is multiplicative. Therefore, the probability that a pair (predictor, predictand) would randomly fall into the category ON-PA, for example is

$$\Pr\{O_2 \cap P_3\} = \Pr\{O_2\}\,\Pr\{P_3\} = \frac{1}{3}\frac{1}{3} = \frac{1}{9}.$$

In more general terms,

$$\Pr\{O_i \cap P_j\} = \Pr\{O_i\}\,\Pr\{P_j\} = \frac{1}{3}\frac{1}{3} = \frac{1}{9}.$$

Hence, if they are independent, a given observation (predictor) can indistinctively be associated with low, normal or high values of the prediction (predictand) variable.

The prediction is termed successful if a low (normal, high) observation matches with a low (normal, high) prediction. Technically:

$$\Pr\{success\} = \Pr\{(O_1 \cap P_1) \cup (O_2 \cap P_2) \cup (O_3 \cap P_3\}$$
$$= \Pr\{O_1 \cap P_1\} + \Pr\{O_2 \cap P_2\} + \Pr\{O_3 \cap P_3\} = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}.$$

Note that the probability of a success corresponds to the total area along the main diagonal.

For negative correlation, the forecast would be successful if a low observation (normal, high) matches the value of high prediction (normal, high),i.e,

$$\Pr\{success\} = \Pr\{(O_1 \cap P_3) \cup (O_2 \cap P_2) \cup (O_3 \cap P_1\}$$
$$= \Pr\{O_1 \cap P_3\} + \Pr\{O_2 \cap P_2\} + \Pr\{O_3 \cap P_1\} = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}.$$

Now the probabilities of success correspond to the area along the secondary diagonal.

3.4 Statistical analysis[17]

3.4.1 Diagnostic statistics

The chi square, χ^2 ,Test

Another measure of discrepancy between the observed (f_{ij}) and expected (e_{ij}) frequencies is proportional to the χ^2statistic given by

$$\chi^2 = \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$, where the total frequency is n.

The χ^2 statistic evaluates how much a contingency table departs from the random case. It can be applied to contingency tables of rectangular contingency. In our case (*3x3*) the contingency is square. The number of degrees of freedom in a table with multiple entries, if the columns and rows are independent variables, is *(M-1)* x *(N-1)*. In our case, the number of degrees of freedom correspond to *(3-1)* x *(3-1) = 4* because the expected frequencies are calculated without recourse to sample estimation of the population parameters. If the *3x3* contingency table is random, the sum of any row or column is *n/3*, and because of that, for any given value of *(i,j),* e_{ij} *= (n/3) (n/3)/ (n) = n /9*. The significance of the χ^2 statistic is determined on the basis of the null hypothesis H_0. If under this hypothesis the calculated χ^2 value is higher than some critical value (such as $\chi^2_{.95}$ or $\chi^2_{.99}$, for significance of 0.05 y 0.01 each), then the observed frequencies differ significantly from the expected frequencies, and H_0 is rejected at the corresponding significance level. Otherwise, H_0 is accepted (not rejected) and the table would be considered to not have predictive value at that level. The statistical significance level of

[17] Some suggestions are included, provided by Dr. Luis Cid of the Statistics Department of the University of Concepción, Chile.

the χ^2 statistic is determined by the cumulative probability starting from *0* to the χ^2 value. Independent values from the rows and columns have significance values that tend towards *0*. When the independent and dependent variables are strongly and positively associated, χ^2 attains high values and the significance tends towards unity (1) at its upper limit. In more graphical terms, Figure 3.2.1 (a) has a lower χ^2 than Figure 3.2.1 (b).

Pearson Coefficient of Correlation

This coefficient is appropriate when both variables have been categorized and has a range of $-1 \le r \le 1$. It is calculated using the following relationships.

$$r = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}$$

Where,

$$ss_r = \sum_i \sum_j f_{ij} (R_i - \overline{R})^2 ,$$

$$ss_c = \sum_i \sum_j f_{ij} (C_i - \overline{C})^2$$

and,

$$ss_{rc} = \sum_i \sum_j f_{ij} (R_i - \overline{R})(C_i - \overline{C}) .$$

R_i and C_j are each the totals from row i, and column j.

The Pearson Correlation Coefficient significance test uses the normalized statistic ***r**** that has an asymptotical normal distribution under the null hypothesis. This statistic is defined as

$$r^* = \frac{r}{\sqrt{\mathrm{var}_0(r)}},$$

where

$$\mathrm{var}_0(r) = \frac{\sum_i \sum_j f_{ij} (R_i - \overline{R})^2 (C_j - \overline{C})^2 - ss_{rc}^2 / n}{ss_r \, ss_c} .$$

This asymptotic variance comes from the multinomial sampling within the framework of a contingency table. It differs from the more familiar form obtained under

the assumption that both variables are continuous and normally distributed[18]. The significance value corresponds to the cumulative probability to the right of the normal distribution. In this way when **r*** has values close to zero (the variables are not correlated), the significance corresponds to 0.5. The significance of variables positively (negatively) correlated tends to 1 (0).

3.4.2 Note on the Cross-Validation[19]

Cross-Validation is a re-sampling technique that operates in a way similar to the bootstrap and permutation tests, dividing the total group of data repeatedly into a subgroup for control and another for verification. The most common situation is that the size of the first group is *n-1* and that of the second is *1*, with *n* different data partitions. The way to apply this technique is as follows[20]:

4) The first value of the dependent variable is removed along with that of the independent variable and the model is recalculated with the *n-1* pairs of remaining data, or the control sample. This is called the reduced model.

5) Using the reduced model calculated in 1), the removed value of the dependent variable is estimated by using the removed value of the independent variable. For comparison purposes this estimated value is divided by the multiple coefficient of determination, R, from the reduced model. These estimated values are called inflated. For example, if the contingency table (of the *n-1* remaining values) and the category of the removed independent observation #1 predict that the removed dependent observation #1 should be in the *below normal* category with a higher probability than in the other categories (*normal or above normal*), then the prediction assigned is *below normal*.

6) The removed data are reincorporated from the dependent and independent variables into the total group of data and the next ones are removed (observation #2), repeating steps 1) and 2), one after another until there are *n* reduced models and *n* estimated data, where *n* is the length of the time series.

Percentages of False Alarms and Detection: Below and Above Normal

The false alarm event occurs when a prediction fails. As an example, consider the situation where the removed dependent observation occupies the normal category where the removed independent value predicts below normal according to the reduced model. In the same manner each one of the *n-1* successive tests are qualified as a success or a false alarm. The diagnostic statistics given by the program correspond to the proportions of success or false alarm (between the *n* tests) for above normal and below normal. Then,

---

[18] Brown, M.B. and Benedetti, J.K. 1977. Sampling behavior of tests for correlation in two way contingency tables. J. Am. Stat. Ass., 72, 309-315.

[19] See: - Ward, N., and C. Folland, 1991. Prediction of seasonal rainfall in the North Nordeste of Brazil using eigenvectors of Sea Surface Temperature. *Int. J. of Climatol.*, **11**, 711-743.

- Wilks, D., 1995. Statistical Methods in the Atmospheric Sciences. Academic Press. 465 pp.

[20] The function crosval.m included in the sourcecodes directory executes this calculation.

these proportions are compared to the expected ones for the case of independent (uncorrelated) variables.

If we use a contingency table as a reference, we calculate the values of *False Alarms* and *Percentages* of *Detection Below* and *Above Normal* as:

$$FARBN = f_{13} / (f_{13} + f_{23} + f_{33}),\ \ FARAN = f_{31} / (f_{11} + f_{21} + f_{31}),$$

$$PODBN = f_{11} / (f_{11} + f_{12} + f_{13}), \text{ and } PODAN = f_{33} / (f_{31} + f_{32} + f_{33}),$$

respectively.

We will now delve more deeply into these concepts for the reader who is interested in the mathematical details; otherwise the user may safely continue to the description of the next statistic.

The probability of associated false alarms for a value under consideration is defined as:

$$\Pr\{\textit{False alarm} \mid O_1\} = \Pr\{P_2 \cup P_3 \mid O_1\} = \Pr\{P_2 \mid O_1\} + \Pr\{P_3 \mid O_1\}$$
$$= \frac{\Pr\{P_2 \cap O_1\}}{\Pr\{O_1\}} + \frac{\Pr\{P_3 \cap O_1\}}{\Pr\{O_1\}} \quad .$$

For independent variables:

$$\Pr\{\textit{False alarm} \mid O_1\} = \frac{1/9}{1/3} + \frac{1/9}{1/3} = \frac{2}{3}.$$

The concept may be simplified by considering the values of the corners of the contingency table within the dominant (most populated) diagonal. In this manner the probability of a false alarm below normal, or *FARBN*, and the probability of a false alarm above normal, or *FARAN*, is defined as:

$$FARBN = \Pr\{P_3 \mid O_1\} = \frac{\Pr\{P_3 \cap O_1\}}{\Pr\{O_1\}}, \textit{and} \; FARAN = \Pr\{P_1 \mid O_3\} = \frac{\Pr\{P_1 \cap O_3\}}{\Pr\{O_3\}},$$

each.

With the corner values of the dominant diagonal we can define the probabilities of a detection below and above normal, *PODAN* and *PODBN*, each. Analogously,

$$PODBN = \Pr\{P_1 \mid O_1\} = \frac{\Pr\{P_1 \cap O_1\}}{\Pr\{O_1\}}, \textit{and} \; PODAN = \Pr\{P_3 \mid O_3\} = \frac{\Pr\{P_3 \cap O_3\}}{\Pr\{O_3\}}.$$

If there is independence *FARBN = FARAN = PODBN = PODAN =1/3*. A strong association between independent and dependent variables lowers the values of *FARBN* and *FARAN* and increases the *PODBN* and *PODAN*.

Hit Rate

As mentioned earlier, the probability of a success or failure corresponds to the probability along the dominant or smallest diagonal, whichever the case. The probability of success in percentages is known as the hit rate (HR). This is calculated as:

$HR = (f_{11} + f_{22} + f_{33}) / n * 100$, for a positive correlation and

where randomly we expect a hit rate $C = 1/3$, that is, a 33.33%.

Skill Score

We can transform the information that the HR coefficient gives us, to construct the Skill Score of SS, as:

$$SS = \frac{HR - 33.33}{100 - 33.33} \times 100.$$

Note that when a prediction is completely random $SS = 0$, and when the correspondence is exact (Pr{Success} = 1) $SS = 100$, that is a perfect group of hits. This last case tells us that the empirical probabilities of the smaller diagonal are null. Negative values of the *SS* indicate that misses dominate in our analysis.

Linear Error in the Probability Space (*LEPS Score*)

Another useful statistic is of the Lineal Error in the Probability Space or *LEPS score*. This statistic is similar to the *SS* with the exception that now the predictions having two erroneous terciles are much more downgraded than those that only have one and we can express it as:

$LEPS = (z_1 / z_2) * 100$,

where z_1 is the sum of the weighted frequencies; for the positive correlation:

$z_1 = 1.35 * f_{11} – 0.15 * f_{12} – 1.20 * f_{13} – 0.15 * f_{21} + 0.30 * f_{22} – 0.15 * f_{23} – 1.20 * f_{31} – 0.15 * f_{32} + 1.35 * f_{33,}$

Now, z_2 is the sum of the weighted frequencies in a perfect group of hits, this means n in our case. Note that if the predictions were random and all the empirical frequencies of the contingency table tended to the same value (z_1) then *LEPS* would tend to 0. On the other hand if we had a perfect prediction the ratio z_1/z_2 would tend to 1 and the *LEPS* would tend to *100*. As with the *SS*, negative values of the *LEPS* would indicate that misses dominate our analysis.

8. Some applications for Climate Forums

4.1 Application for the climate forums I

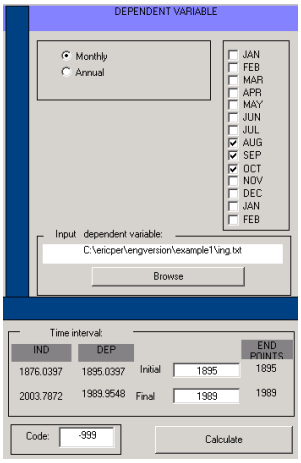
For this first example we will use the precipitation data from the Ingenio San Antonio station (87.05° W, 12.53° N, 35 msnm), located on the Pacific coast of

Nicaragua, with monthly records from 1895 to 1989. The data are located in the directory `...\ejemplo1\ing.txt`, and the code for the missing data is `-999`. The file `ing.txt` contains two columns, the first is the year plus the mid-month fraction (time vector) and the second is the accumulated monthly precipitation in mm for the Ingenio San Antonio station:

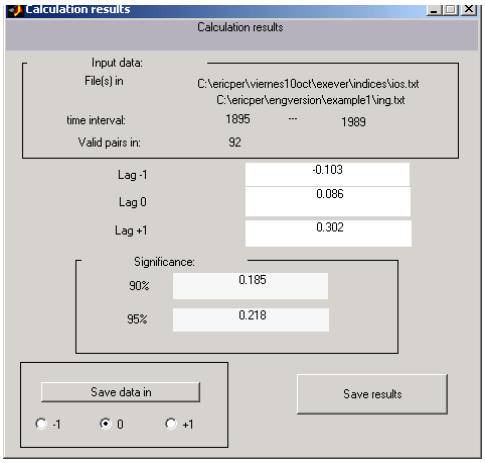
```
%      time          Precip.-Ingenio San Antonio
    1895.039726            0.000000
    1895.123058            0.000000
    1895.206390            0.000000
    1895.289722            0.000000
         ⋮                     ⋮
    1989.704799            664.000000
    1989.788131            252.000000
    1989.871463            41.000000
    1989.954795            4.000000
```

In this example we will explore the relationship between the precipitation of ASO at Ing. San Antonio and the Southern Oscillation Index or SOI for the periods JFMA, MJJ, ASO, ND. As a first step when executing the program `wfcc.exe` or `wfcceng.exe`, we must click on the button `Internal` under independent variable, because the SOI is one of the climate indices provided with the program, then for the months we mark JAN, FEB, MAR y APR, then we click on load.

Next, under dependent variable we click on `monthly`, because the `ing.txt` file contains monthly data, then in the months we mark AUG, SEP and OCT, then we look for the `ing.txt` file in the corresponding directory and click on `open`. We then choose the period to use, for our example 1895 under `Initial` and 1989 under `Final`, and we must also type in the code for the missing data, `-999`.

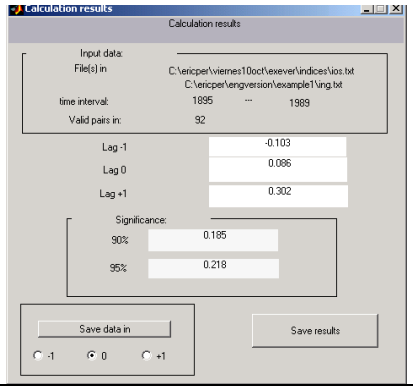


Once this is done we click on `Calculate`, which will deploy the following screen:

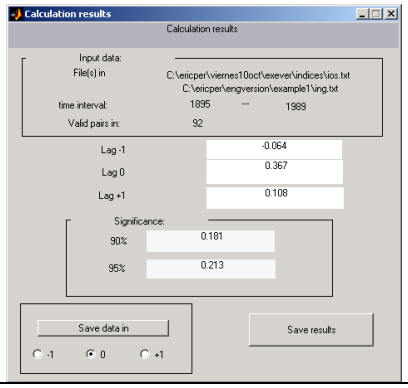


These results refer to the cross correlation function or CCF between the SOI for JFMA of the year before and the precipitation at Ing. San Antonio during ASO, or lag –1, between the SOI for JFMA of the same year and Ing. San Antonio during ASO, or lag 0, and between the SOI for JFMA of the next year and Ing. San Antonio during ASO, or lag +1. The 90 and 95% significance levels of these correlation values are also included. Repeating the steps mentioned before but selecting MJJ, ASO and ND as periods for the independent variable we obtain the next results:

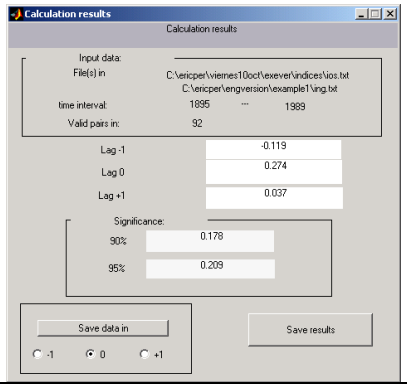
SOI-JFMA



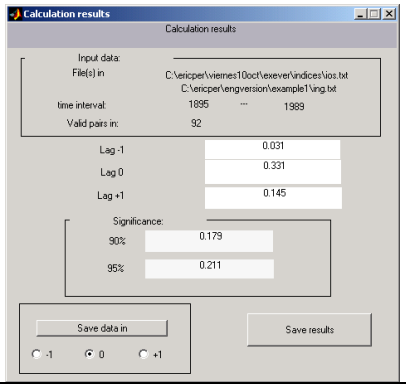
SOI-MJJ



SOI-ASO



SOI-ND

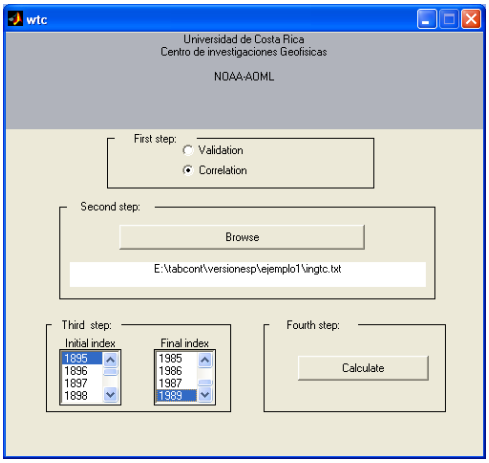


Note that the value of the CCF between the SOI – MJJ and the precipitation of ASO from lag 0 is 0.367, significant at 95%, thus at first sight we could use a classic prediction scheme. To proceed this way we click on the Save data in data button on the screen of the

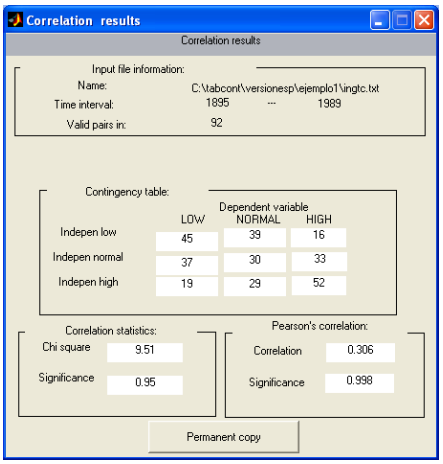
results from SOI-MJJ and we save them in the ...\example\ingios.txt file. The file ingios.txt contains three columns, the first one is the year, the second one is the average of the average from SOI for the trimester MJJ and the third the average of anomalies of the precipitation in the Ing. San Antonio station for the trimester ASO. The lines preceded by the % symbol are taken as comments and are ignored in subsequent calculations. Since we cannot include the contingency analysis of the missing data, we proceed to put the % symbol in front of the lines that contain -999 or NaN in any of its columns, using any convenient text editor:

```
% File with annual series
% built from the following source files:
% Input file C:\ericper\viernes10oct\exever\indices\ios.txt
% Input file C:\ericper\engversion\example1\ing.txt
% Initial year: 1895.00
% Final year: 1989.00
  1.8950e+003     -4.4333e+000       4.1033e+002
  1.8960e+003     -3.1133e+001       2.4833e+002
  1.8970e+003     -6.3333e+000       4.7300e+002
  1.8980e+003      6.3333e-001       3.3133e+002
  1.8990e+003     -7.8000e+000       3.5400e+002
  1.9000e+003      9.5667e+000       3.8067e+002
% 1.9010e+003      1.1300e+001              NaN
% 1.9020e+003      3.9333e+000              NaN
  1.9030e+003      4.3667e+000       3.2567e+002
```

After saving this file we run the program wtc.exe or wtceng.exe. First we click the Correlation button, then we look for the file: ingSOI.txt and select 1895 for the Initial Index and 1989 for the Final Index.



Finally we click on Calculate, which deploys the next screen:



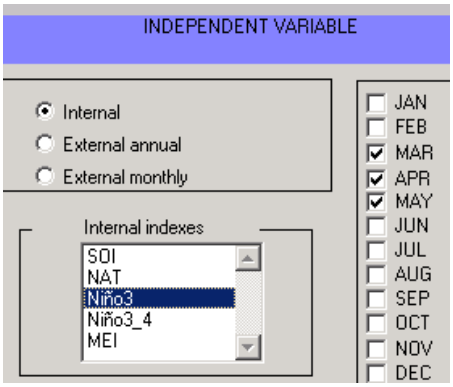
If we suppose that we will observe a negative SOI value during MJJ (in the lowest tercile), our suggestion for the climate forum on the expected probabilities for the precipitation at Ing. San Antonio during ASO will be: 45% BN, 39% DN y 16% AN.

4.2 Application for the climate forums II

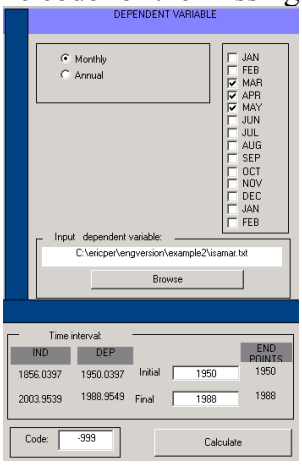
For this second example we will use the precipitation data of the Isabel María station (79.56° W, 1.83° N, 4 msnm), from Ecuador, with monthly records from 1950 to 1988. The data are found in the directory `...\example2\isamar.txt`, and the code for the missing data is `-999`. The file `isamar.txt` contains two columns, the first one is the year plus the mid-month fraction (time vector) and the second one is the accumulated monthly precipitation in mm for the Isabel María station:

```
%       time           Pcp.-Isabel Maria
  1.9500397e+003  3.6910000e+002
  1.9501231e+003  4.9310000e+002
  1.9502064e+003  3.1700000e+002
  1.9502897e+003  1.6760000e+002
        ⋮                      ⋮
  1.9887049e+003  -9.9900000e+002
  1.9887883e+003  -9.9900000e+002
  1.9888716e+003  -9.9900000e+002
  1.9889549e+003  -9.9900000e+002
```

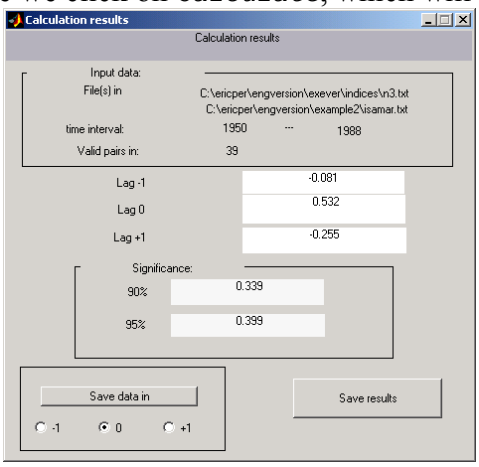
In this example we will explore the relationship between the precipitation during MAM at Isabel María and the Niño 3 index for the DJF, MAM, JJA, SON periods. As a first step when executing the `wfcc.exe` or `wfcceng.exe` program, we must click on the `Internal` button for the independent variable, because Niño 3 is one of the climate indices provided with the program, then for the months we select MAR, APR and MAY and we finish by clicking on `Browse`.



Next, under `dependent variable`, we select `monthly`, because the file `isamar.txt` contains monthly data, then for the months we select MAR, APR and MAY, after that we search for the `isamar.txt` file in the corresponding directory and click on `open`. Following this we select the period to use, in our example, 1950 as `Initial` and 1988 as `Final`, and we type the code for the missing data, `-999`.



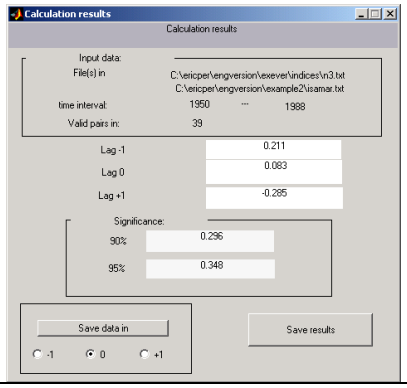
Once this is done we click on `Calculate`, which will deploy the following screen:



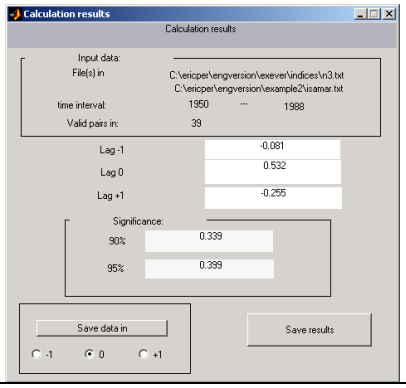
These results refer to the Cross Correlation Function or CCF between Niño 3 for MAM of the year before and the precipitation at Isabel María during MAM, or lag -1, between Niño 3 for MAM of the same year and the precipitation at Isabel María during MAM, or lag 0, and between Niño 3 for MAM of the year following the precipitation and Isabel María during MAM, or lag +1. The 90 and 95% significance levels of these

correlation values are also included. Repeating the steps done before but selecting DJF, JJA and SON as periods for the independent variables we obtain the following results:

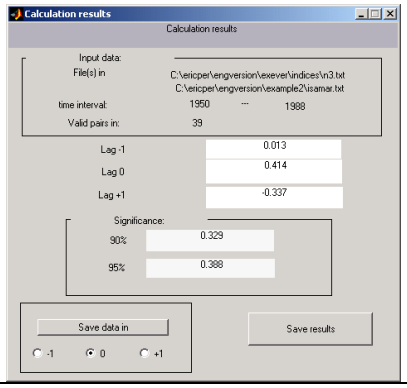
Niño 3-DJF



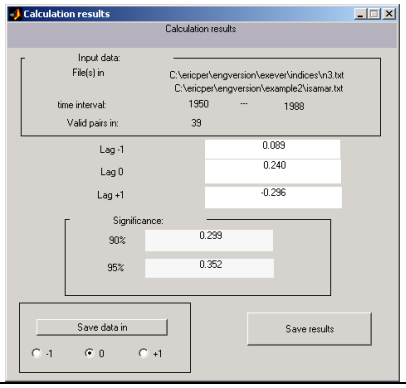
Niño 3-MAM



Niño 3-JJA



Niño 3-SON

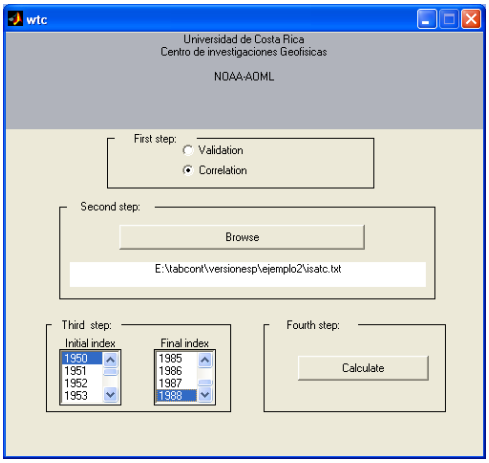


Note that the CCF value for Niño 3-MAM versus the precipitation during MAM of lag 0 is 0.53, significant at 95%. To use this result for climate prediction we must adopt the scheme known as "perfect prognosis" based on the results of a predicted scenario for the Niño 3 index, given that the relationship of Niño 3 to the precipitation at Isabel María is a contemporaneous one (lag 0). To do this we click on the `Save data in` button on the results screen for Niño 3-MAM and we save them in the `...\example2\isamarn3.txt` file. The `isamarn3.txt` file contains three columns, the first is the year, the second is the average of the values of Niño 3 for the MAM trimester and the third is the average of the precipitation anomalies at Isabel María during the MAM trimester. The lines preceded by the % symbol are comments. Since the missing data cannot be included in the contingency analysis, we proceed to put the % symbol in front of the lines that contain `-999` or `NAN` in any of its columns, using any convenient text editor:

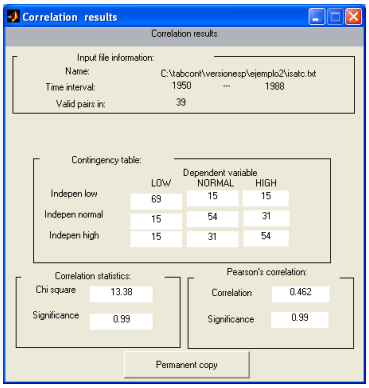
```
% Annual Series file:
% Composed by the next format files:
% Redden file C:\ericper\viernes10oct\exever\indices\n3.txt
% Redden file C:\ericper\foroguaya\ejemplo2\isamar.txt
% Initial year: 1950.00
% Final year: 1988.00
% Year             Niño 3-MAM(0)     Pcp.Isabel Maria-MAM
  1.9500e+003   -7.7667e-001    1.6187e+002
  1.9510e+003   -8.3333e-002    2.3890e+002
```

```
1.9520e+003   -1.3333e-001    1.2887e+002
1.9530e+003    5.0667e-001    4.3687e+002
1.9540e+003   -7.4000e-001    1.3697e+002
1.9550e+003   -9.3333e-001    3.0887e+002
1.9560e+003   -5.0667e-001    2.4790e+002
1.9570e+003    4.2667e-001    4.4293e+002
1.9580e+003    3.3000e-001    3.3827e+002
```

After saving this file we execute the `wtceng.exe` or `wtc.exe` file. First we select the `Correlation` button, then we look for the file `isamarn3.txt` and select 1950 as `Initial Index` and 1988 as `Final Index`:



Finally we click on `Calculate`, which will deploy the following screen:



If we expect Niño 3 to be normal during MAM (middle tercile), our suggestion for the climate forum on the probabilities expected for the precipitation at Isabel María during MAM will be: 15% BN, 54% DN y 31% AN.

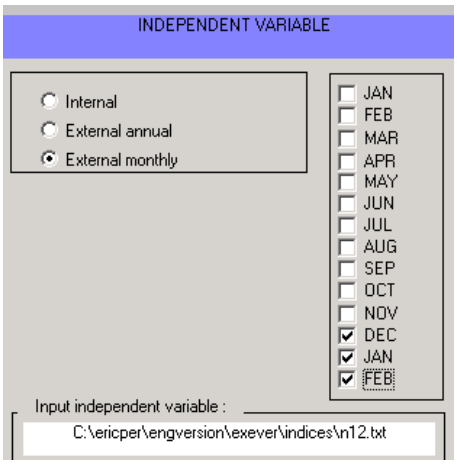
4.3 Application for the climate forums III

For this third example we will use the precipitation data of the Phillip Goldson station (88.30° W, 17.53° N, 5 msnm), located in the international airport of Belize, with monthly records from 1960 to 2002. The data are located in the `...\example3\phi.txt`, directory and do not contain missing data. The `phi.txt` file contains two columns, the

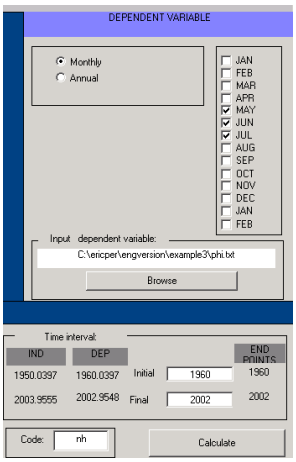
first is the year plus the mid-month fraction (time vector) and the second the accumulated monthly precipitation in mm for Phillip Goldson:

```
%Year        Pcp.Phillip Goldson
1960.039726          68.100000
1960.123056          20.300000
1960.206386          30.700000
1960.289717          95.000000
    ⋮                      ⋮
2002.704804         106.400000
2002.788134          35.500000
2002.871464          97.800000
2002.954795          244.100000
```

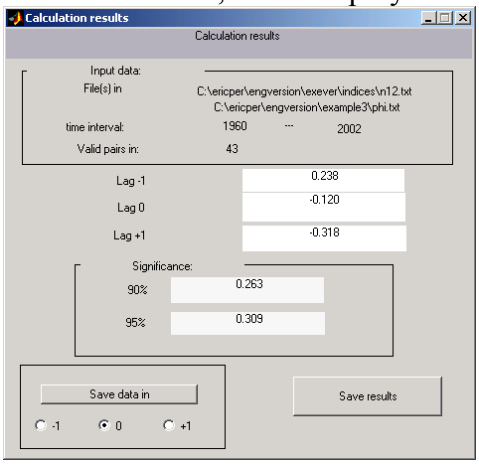
In this example we will explore the relationship between the precipitation during MJJ at Phillip Goldson and the Niño 1+2 Index for the DJF, MAM, JJA, SON periods. As a first step when executing the `wfcceng.exe` or `wfcc.exe` program, we must click on the `monthly external` button, under independent variable, because the Niño 1+2 is not one of the climate indices provided with the program. Then for the months we select DEC, JAN and FEB, followed by a click on `Browse` to find the `n12.txt` file that contains the monthly data of the Niño 1+2 index and finally we click on `open`.



Next, under `dependent variable`, we select `monthly` because the `phi.txt` file contains monthly data, then for the months we select MAY, JUN and JUL and we look for the `phi.txt` file in the corresponding directory and click on `open`. Then we select the period to be used, for our example 1960 as `Initial` and 2002 as `Final`, and we type the code for the missing data, in this case we can specify `nh` because we do not have missing data.

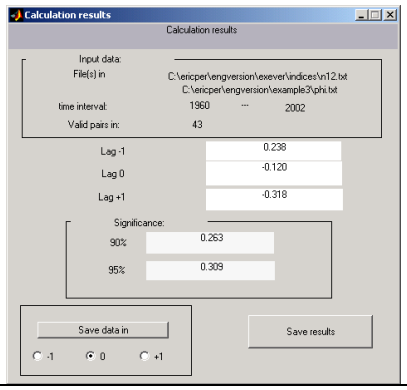


Once this is done we click in `Calculate`, which deploys the following screen:

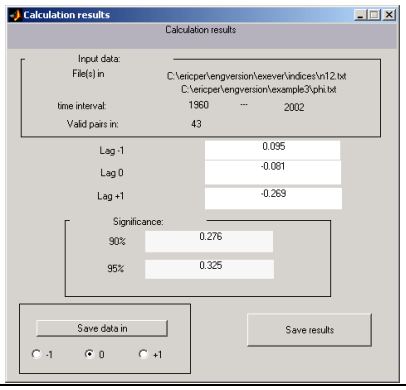


These results refer to the Cross Correlation Function or CCF between Niño 1+2 for DJF of the year before and the precipitation at Phillip Goldson during MJJ, or lag -1, between Niño 1+2 for DJF of the same year and the precipitation at Phillip Goldson during MJJ, or lag 0, and between Niño 1+2 for DJF of year following and precipitation at Phillip Goldson during MJJ, or lag +1,where Dec is the month that defines the years -1, 0 and +1. The 90 and 95% significance levels for these correlation values at are also included. Repeating the steps given before but selecting MAM, JJA and SON as periods for the independent variable we obtain the following results:

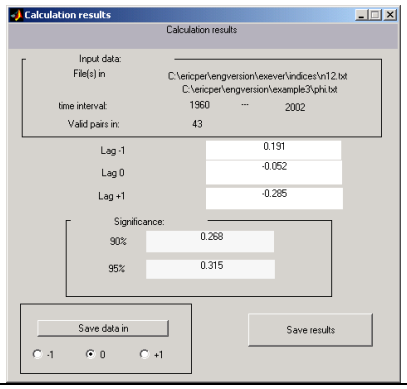
Niño 1+2-DJF



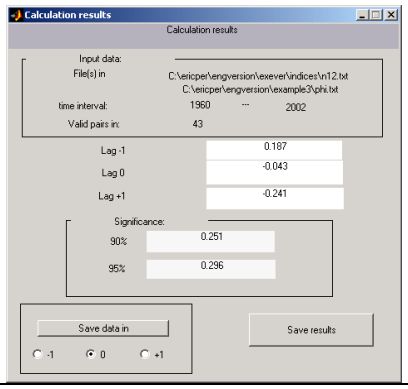
Niño 1+2-MAM



Niño 1+2-JJA



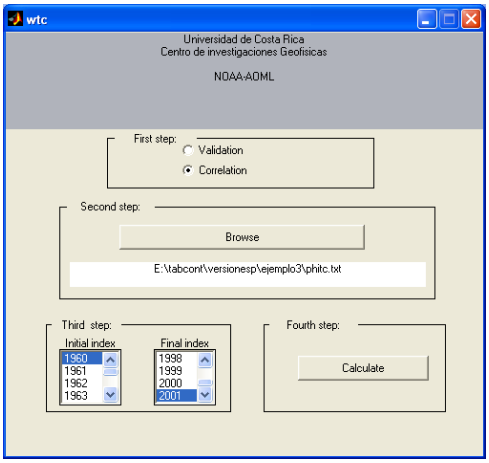
Niño 1+2-SON



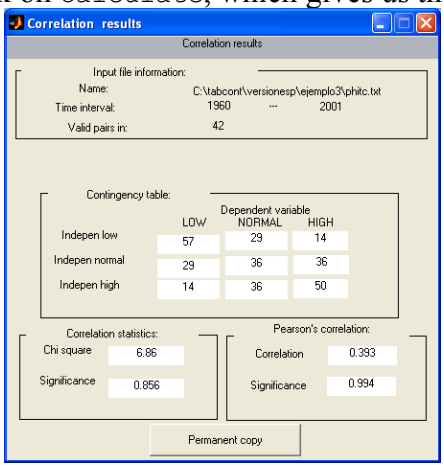
Note that the value of the CCF between the Niño1+2-DEF of year -1 and the precipitation during MJJ is 0.24, which is not significant at 90%. Nevertheless, we find other high values of the correlation for lag +1, which tells us that precipitation precedes the climate index. This is difficult to use in a predictive scheme; on the other hand we must remember that we are using the results of the CCF as a guide for the analysis of the contingency table and not to adjust a linear regression model. Next, we click on the `Save data in` button on the results screen for Niño1+2-DJF but do click also in –1 and we save them in the `...\example3\phin12.txt` file. The `phin12.txt` file contains three columns, the first one is the year, the second is the average of the Niño 1+2 values for the DJF trimester and the third one the anomalies of the precipitation at the Phillip Goldson station for the MJJ trimester. The lines preceded by the symbol % are comments:

```
% File with annual series
% built from the following source files:
% Input file C:\ericper\engversion\exever\indices\n12.txt
% Input file C:\ericper\engversion\example3\phi.txt
% Initial year: 1960.00
% Final year: 2002.00
  1.9600e+003      1.2333e-001       3.2227e+002
  1.9610e+003     -5.4667e-001       2.9987e+002
  1.9620e+003     -6.9000e-001       1.0563e+002
  1.9630e+003     -4.6333e-001       1.9810e+002
  1.9640e+003     -4.6333e-001       1.3870e+002
  1.9650e+003      4.2333e-001       3.0113e+002
  1.9660e+003     -5.4667e-001       1.4970e+002
  1.9670e+003     -1.3700e+000       1.7667e+002
  1.9680e+003     -2.3333e-002       2.4160e+002
```

After saving this file we execute the `wtc.exe` or `wtceng.exe` program. First we click on the `Correlation` button, next we look for the `phin12.txt` file, then we select 1960 as `Initial Index` and 2001 as `Final Index`:



In last step we click on `Calculate`, which gives us the following screen:



As part of the example, if we observed Niño 1+2 above normal during DJF (third tercile), our suggestion for the weather forum on the expected probabilities in the precipitation of Phillip Goldson during MJJ, will be: 14% BN, 36% DN y 50% AN.

4.4 Application for the climate forums IV

As our dependent variable we will now analyze the behavior of an index of the surface air temperature for Central America[21] averaged for the months of May, June and July (MJJ) from 1958 to 1998. This index is representative of the behavior of this atmospheric parameter over most of the region as shown in Fig. 4.4.1.

[21] The index was built using the analysis of principal components. Obtained data as a collaboration of the CRN073-IAI project.

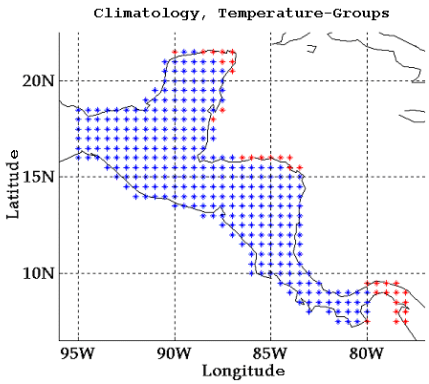
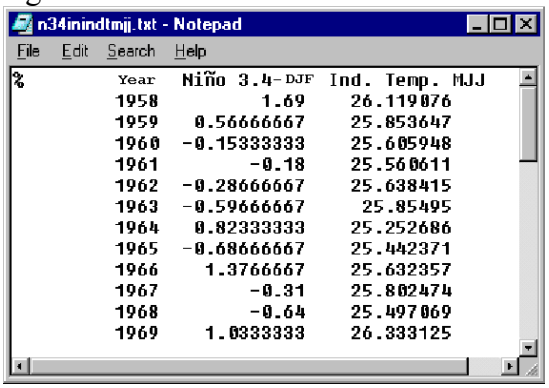


Fig. 4.4.1. The blue dots represent the locations where the index of the surface air temperature used in this example dominates the variance[22]. This EOF explains around 80% of the variance in the region.

The files to use with the `wfcc.exe` or `wfcceng.exe` program, contained in the `...\example4` directory are:

`N34inindtmjj.txt`
`N34prindtmjj.txt`
`N34veindtmjj.txt`
`N34otindtmjj.txt`,

where `N34` corresponds to the index of sea surface temperature for the Niño 3.4 region; `in` is winter (DJF), `pr` is spring (MAM), `ve` is summer (JJA) and `ot` is fall (SON), all for the northern hemisphere. The first column of these files is the year, the second is the corresponding value of Niño 3.4 and the third is the air temperature index for MJJ.

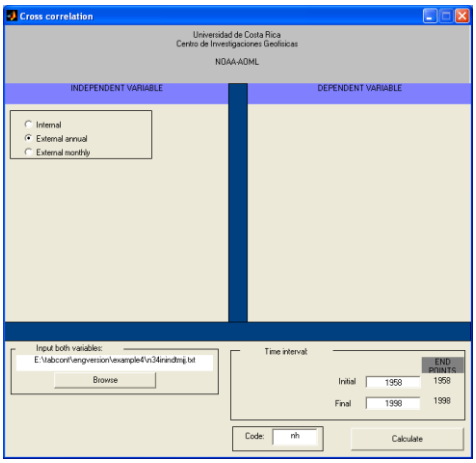

n34inindtmjj.txt - Notepad

File Edit Search Help

% Year	Niño 3.4-DJF	Ind. Temp. MJJ
1958	1.69	26.119076
1959	0.56666667	25.853647
1960	-0.15333333	25.605948
1961	-0.18	25.560611
1962	-0.28666667	25.638415
1963	-0.59666667	25.85495
1964	0.82333333	25.252686
1965	-0.68666667	25.442371
1966	1.3766667	25.632357
1967	-0.31	25.802474
1968	-0.64	25.497069
1969	1.0333333	26.333125

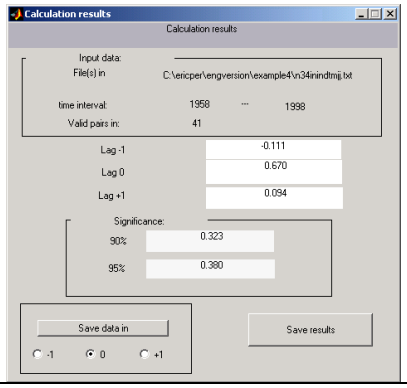
These files have the three columns in the correct order to estimate the CCF, then as a first step when executing the `wfcc.exe` or `wfcceng.exe` file we must click the `Annual External` button under `independent variable`, then a click on `Browse` to find the corresponding files and finally click on `open`. After this we choose the period to be used, for our example 1958 as `Initial` and 1998 as `Final`, and we type the code for the missing data, in this case `nh` because there are no missing data.

- [22] See: Alfaro, E., 2000: Response of Air Surface Temperatures over Central America to Oceanic Climate Variability Indices. *Tópicos Meteorológicos y Oceanográficos*, 7(2), 63-72.

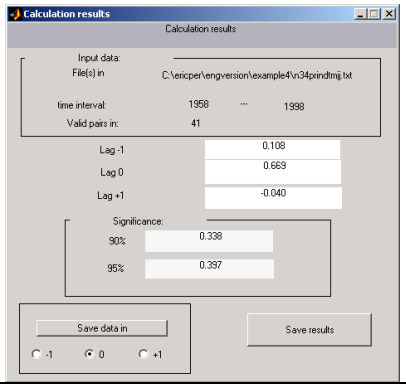


Once this is done we click on `Calculate`, and repeat the process for the four files corresponding to the four seasons of the year for the Niño 3.4 index, then we obtain the following:

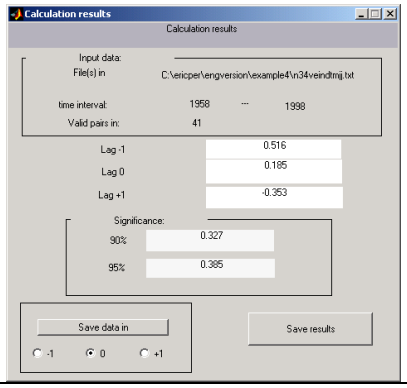
Niño 3.4-DJF



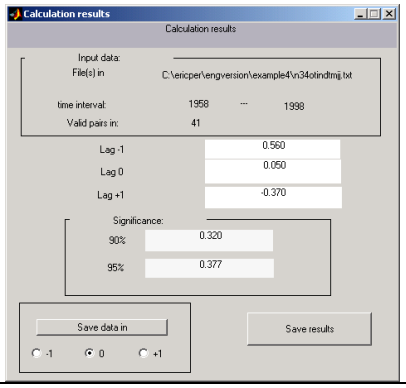
Niño 3.4-MAM



Niño 3.4-JJA

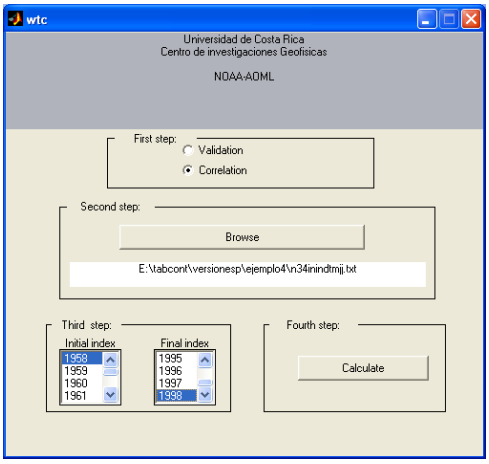


Niño 3.4-SON

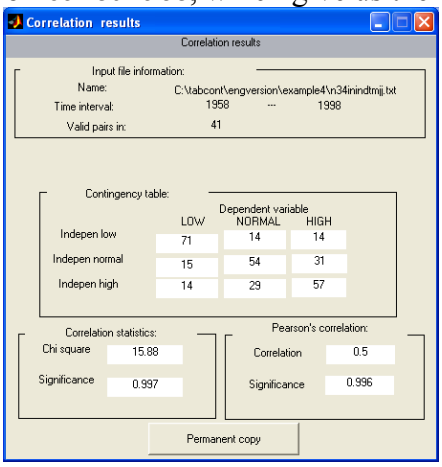


These results show that the largest correlation is 0.67 for the year 0 winter of the Niño 3.4 index versus the temperature index for MJJ, significant at 95% (Note: in this particular example the year 0 for DJF was defined by Jan and Feb instead Dec).

We next execute `wtc.exe` or `wtceng.exe`. First we click the `Correlation` button, then we open the `N34inindtmjj.txt` file and select 1958 as `Initial Index` and 1998 as `Final Index`:



Finally we click on `Calculate`, which give us the next screen:



These results suggest a classic prediction scheme because the independent variable, Niño 3.4-DJF, precedes the dependent variable, the MJJ air temperature index. Hence, for the scenario of the independent variable we simply select the Niño 3.4 tercile currently observed for DJF. Thus if we observe that the Niño 3.4 during the past DJF was above normal (upper tercile), our suggestion for the surface air temperature during MJJ at the climate forum would be: 14% BN, 29% DN y 57% AN.

9. Contact information for the authors

Dr. F. Javier Soley (programming and statistical principles)
CIGEFI-Escuela de Fisica
Universidad de Costa Rica
2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320

Fax: (506) 234-2703
Email: fjsoley@racsa.co.cr

Dr. Eric J. Alfaro (statistical principles and practical applications in the RCOFs)
CIGEFI- Escuela de Fisica
Universidad de Costa Rica
2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320
Fax: (506) 234-2703
Email: ejalfaro@cariari.ucr.ac.cr, ealfaro@cosmos.ucr.ac.cr

Dr. David B. Enfield (organization and applications practical applications in the RCOFs)
NOAA.AOML/PhOD
4301 Rickenbacker Causeway
Miami, Fl 33149
Email: David.Enfield@noaa.gov

The conversion of the functions to C++ and the programming of the graphical user interfaces were done by an electrical engineering student from the UCR:

Vilma Aguilar
CIGEFI
Universidad de Costa Rica
2060-Ciudad Universitaria Rodrigo Facio
San José, Costa Rica
Tel: (506) 207-5320
Fax: (506) 234-2703
Email: vilma@scratchy.emate.ucr.ac.cr

General Warning. All versions of the software (See Apendix) provide extensive diagnostic statistics designed to help the user to discern the true significance of the relationship between predictor and predictand. Ultimately, the decision as to how much credence to give to a result depends on this judgment. Users who have a greater command of the underlying statistical principles are less likely to commit errors of judgment or make unjustified projections. The user is advised to study the manual and examples carefully in regard to the diagnostic statistics and to make maximum use of them.

6. Acknowledgments

This work was developed thanks to the support of the NOAA-OGP by the project *A special proposal to improve regional climate Outlooks in Latin America* and the University of Costa Rica through the projects VI-112-99-305, ED-1040 and MM5-UCR-CRRH.

Appendix

Description of the evolution of the contingency table programs

UPDATE HISTORY

1-) April - 2003

Two programs: corcruz3.exe and valdos.exe. They run from the DOS prompt. They were first presented and distributed during the COF at San Pedro Sula, Honduras, April, 2003. No outright errors or malfunctionings (bugs) have been found in this version. This version was/is available only in Spanish and only for use with DOS (no graphical interface).

Corcruz3.exe calculates three lags (0, ±1) of the cross correlation function and the large lag standard error at the significance levels of 90 and 95 %. The inputs are ASCII files with three columns containing a time index, an independent annual variable and a dependent annual variable.

Valdos.exe calculates the 3x3 contingency table for the independent annual variable and the dependent annual variable. There are two options, "correlación" and "validación". The first one is used for negative or positive correlation between the variables and the second one for cross validation. The inputs are similar to those of corcruz3.exe.

Statistics calculated for the "validación" option

h)Hit rate
i)Skill Score
j)False alarms below and above normal
k)Detection probability above and below normal
e) Pearson's correlation and its significance
h)Chi square test and its significance
i)G square test and its significance
j)Linear error in probability space

Statistics calculated for the "correlación" option

a)Pearson's correlation and its significance
b)Chi square test and its significance
c)G square test and its significance

Limitations: some users, mainly those who have only used the later versions of MS Windows have difficulties because they are not familiar with DOS.

2-) November - 2003.

First windows environment version presented in Guayaquil's COF, November, 2003. The programs were called wfcc.exe y wtc.exe. These are the successors of corcruz3.exe and valdos.exe, respectively. At this point the software was still available only in the Spanish version.

Wtc.exe has the same functions as valdos.exe, but calculates only chi square if there are values near zero in the contingency table or G square if there are no values near zero. It also calculates the appropriate test's significance.

There is one known bug: if the cancel button is pressed in the file name window, the program aborts and must be re-launched.

Several functions were added to wfcc.exe. It accepts annual input as corcruz3.exe does but also accepts monthly time series for the independent and dependent variables as inputs. In this way the user is saved the extra step of separately calculating annualized data from monthly data. It also includes some of the most used climate variability indices as default (non-user supplied) choices. The user specifies the consecutive months to be used as an averaged season to calculate the annual time series. It calculates the same statistics as corcruz3.exe. It is possible to save in an ASCII file the data used in the calculation for lag 0.

Known bugs:

e)If the cancel button is pressed in the file name window, the program aborts.
f)The waiting bar does not progress and gives warning messages.
g)Does not handle well the missing data in the monthly variable "independiente" and "dependiente options".
h)When months that span two years are selected, the cross correlation has an extra bias.
i)Shows as data pairs read the number of total lines in the input file.

3-) January -2004

This is the first version available in English as well as Spanish and was prepared for the English-speaking Caribbean in anticipation of the COF in Kingston, Jamaica in March 2004. It is almost the same as the November - 2003 version, but with the

following bugs corrected for the wfcc.exe (or wfcceng.exe in the English version) program:

e)Now handles correctly any missing data in the monthly independent variable ("variable independiente") and dependent ("dependiente") options.
f)If months that span two years are selected in the monthly variable options, the cross correlation is now calculated without an extra bias.
g)Now shows correctly the data pairs read.
h)The *progress bar's execution* (wait bar) does not give warning messages.

The following bugs persist:

f)If the cancel button is pressed in the file name window, the program aborts.
j)The progress bar does not work.

This program can also save the data used in the cross correlation calculation for the -1, 0 or +1 lag. The saved output can be used directly by wtc.exe (or wtceng.exe in the English version).

Wtceng.exe did not change.

4-) December, 2004-January, 2005.

Wfcceng.exe did not change.

For wtceng.exe, the calculation and screen output of chi2 and its significance was corrected. A minor error in the rounding of some neutral probabilities was detected and corrected. It occurred mainly when with few data or when the data vector had the same value repeated many times, e.g. dry season. Notice that in these cases the contingency analysis is not recommended at all. When the pairs of data used are less than 45 it displays a warning message in the DOS screen. There is also now a DOS output for the contingency table with the absolute frequencies which some users find useful. The significance of the Pearson correlation statistics is now calculated by the more conservative method of large-lag standard error, which accounts for any serial correlation in the data. This will give a more faithful estimate of the true significance, but one that is generally lower than before.