

**Evaluación de dos métodos de estabilidad
fenotípica a través de validación
cruzada**

JAIRO ALBERTO RUEDA RESTREPO

Ingeniero Forestal

Director: JOSÉ MIGUEL COTES TORRES

Agrónomo Ph.D.

**FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA**

UNIVERSIDAD NACIONAL DE COLOMBIA

SEDE MEDELLÍN

2006

CONTENIDO

	Página
Introducción	7
Marco Teórico	9
Modelo de Eberhart y Russell	9
Modelo de componentes de varianza de Shukla	11
La técnica de validación cruzada	16
El estimador <i>Jackknife</i>	17
Metodología	22
<hr/>	
Resultados y discusión	27
Referencias	34
Anexo	

LISTA DE TABLAS

- Tabla 1.** Significado de los parámetros de estabilidad obtenidos por la metodología de Eberhart y Russell....10
- Tabla 2.** Estadísticos básicos para el valor del RDMSP al realizar 25000 iteraciones del proceso de validación cruzada sobre los datos de trigo.....26
- Tabla 3.** Pendiente de la recta estimada por el estadístico de Eberhart & Russell, con base en 25000 remuestros sin reemplazo.....27
- Tabla 4.** Sigma-i para cada genotipo estimado por el estadístico de Eberhart & Russell con base en 25000 remuestros sin reemplazo.....30
- Tabla 5.** Estimación de intervalos de confianza al 95% para la varianza de Shukla, con base en 25000 remuestros sin reemplazo.....31

LISTA DE FIGURAS

- Figura 1.** Intervalos de confianza al 95% para la estabilidad genotípica de nueve genotipos de trigo ensayados en 20 localidades.....27
- Figura 2.** Estadístico RDMSP para el estadístico de Eberhart y Russell.....29
- Figura 3.** Estadístico RDMSP para el estadístico de Shukla..29
- Figura 4.** Intervalos de confianza al 95% para Sigma en cada genotipo en la estimación de Eberhart y Russell, con base en 25000 remuestros sin reemplazo.....30
- Figura 5.** Intervalos de confianza al 95% para la varianza de Shukla, con base en 25000 remuestros sin reemplazo.....31

LISTA DE ANEXOS

[Anexo](#). Código en SAS para la validación cruzada.
Teste.sas y pgm.sas.

Evaluación de dos métodos de estabilidad fenotípica a través de validación cruzada

JAIRO ALBERTO RUEDA RESTREPO¹

Director: JOSÉ MIGUEL COTES TORRES²

Mayo de 2006

ABSTRACT Una de las principales preocupaciones de los mejoradores de plantas es la evaluación de la estabilidad fenotípica mediante la realización de pruebas regionales o multiambiente. Existen numerosos métodos propuestos para el análisis de estas pruebas regionales y la estimación de la estabilidad fenotípica. En este trabajo se compara el método de regresión propuesto por Eberhart & Russell y el de componentes de varianza propuesto por Shukla, siguiendo un esquema de validación cruzada, utilizando como indicador de un mejor modelo el parámetro RDMSP (*Root of mean squared error of prediction*). Adicionalmente, se calculó la cantidad de veces que un modelo fue mejor que el otro, en el esquema de validación cruzada. Fueron utilizados datos provenientes de 20 pruebas multiambiente de maíz, cada una con nueve genotipos, plantadas bajo un DBCA (diseño en bloques completos al azar) con cuatro repeticiones. Se encontró que, según el RDMSP, el mejor modelo para predecir el rendimiento futuro de un genotipo es el método de Eberhart & Russell, lo cual ocurrió el 90.6% de la veces evaluadas.

Palabras Clave - KEY WORDS: Interacción genotipo - ambiente, adaptabilidad fenotípica, Varianza de Shukla, Pruebas regionales.

1) Instructor Asociados al Departamento de ciencias Agronómicas de la Facultad de Ciencias Agropecuarias. El trabajo se presenta como requisito para optar al título de especialista en estadística, Facultad de Ciencias. U.N. jarueda@unal.edu.co.

2) Profesor Asistente del Departamento de Ciencias Agronómicas, Universidad Nacional de Colombia, Medellín, Colombia.

Introducción

Actualmente existen muchas metodologías de análisis de estabilidad fenotípica. El método tradicional consiste en el análisis conjunto de los experimentos, considerando todos los ambientes y el posterior desdoblamiento de la suma de cuadrados de efectos de los ambientes y de la interacción genotipo-ambiente (GxA), en efectos de ambientes dentro de cada genotipo. La variación de ambientes dentro de cada genotipo es utilizada como un estimado de la estabilidad, de tal forma que el genotipo que tenga un menor cuadrado medio en los ambientes, será considerado el más estable (Cruz & Regazzi 2001).

Los estudios de la interacción genotipo-ambiente proveen información sobre el comportamiento de cada genotipo frente a variaciones ambientales. Por esto, se realizan análisis de estabilidad y adaptabilidad fenotípica, mediante los cuales es posible identificar cultivares de comportamiento relativamente previsible. En este trabajo, la estabilidad fenotípica se estima por los métodos de Shukla (1972) y de Eberhart & Russell (1966) sobre los rendimientos de nueve variedades de maíz evaluados en 20 ambientes, en pruebas internacionales realizadas por el CIMMIT (Centro de mejoramiento de maíz y trigo).

La base de datos fue dividida usando un esquema de muestreo aleatorio que extraía, de cada celda de la

tabla de rendimientos genotipo x ambiente, una replica. Con la información modificada de esta manera, se calcularon de nuevo los estadísticos de estabilidad por los métodos Shukla y Eberhart & Russell. Con los percentiles 2.5 y 97.5 se consigue un intervalo de confianza tipo Jackknife para los parámetros de estabilidad, también se compara la eficiencia relativa de los dos métodos con la media cuadrática de los desvíos.

Con los dos modelos se estima el rendimiento para cada genotipo y ambiente y la diferencia en la estimación (que excluye la i-ésima observación) con respecto a la que arroja usando el conjunto completo de observaciones, que se asume como la verdadera.

El proceso se realizó 25000 veces, con nueve genotipos y 20 localidades. Para tal efecto se implementaron rutinas de computador (en SAS) que permiten realizar el proceso de forma aleatoria, pero controlando no repetir alguna combinación extraída.

Marco Teórico

Modelo de Eberhart y Russell

Los métodos propuestos por Finlay & Wilkinson (1963) y mejorado por Eberhart & Russell (1966) se basan en el análisis de regresión lineal, que mide la respuesta de cada genotipo con la variaciones ambientales.

La regresión de Eberhart & Russell (E&R) difiere de la regresión de Finlay - Wilkinson en que las desviaciones d_{ij} tienen una varianza separada para cada genotipo (Piepho 1999). Estas varianzas pueden ser denotadas como: $\sigma^2_{d(i)}$ y $\sigma^2_{f(i)}$ (Piepho 1999). Los parámetros de estabilidad se proveen por la regresión sobre un índice ambiental para cada variedad (genotipo) una función de desviaciones al cuadrado. Estos parámetros son definidos por el modelo (Eberhart & Russell 1966):

$$\bar{Y}_{ij} = \mu_i + \beta_i I_j + \bar{e}_{ij} \quad [1]$$

Con $i=1, \dots, v$ (genotipos); $j=1, \dots, n$ (ambientes); μ_i la media de cada cultivar y β_i los coeficientes de regresión que miden la respuesta de cada genotipo al cambio en índice ambiental; I_j es el índice ambiental obtenido como el promedio de todas las variedades en el j -ésimo ambiente menos la gran media o media general (Eberhart & Russell 1966)

$$I_j = \frac{\sum_j y_{ij}}{v} - \frac{\sum_i \sum_j y_{ij}}{vn} \quad [2]$$

Con $\sum_j I_j = 0$.

El primer parámetro de estabilidad es el coeficiente de regresión de forma usual:

$$b_i = \frac{\sum_j Y_{ij} I_j}{\sum_j I_j^2} \quad [3]$$

La suma de las desviaciones -valor observado respecto del valor estimado- proveen una aproximación al parámetro de estabilidad fenotípica dado por (Eberhart & Russell 1966):

$$S_{d_i}^2 = \frac{\sum_j \hat{\delta}_{ij}^2}{(n-2)} - \frac{S_e^2}{r} \quad [4]$$

donde $\frac{S_e^2}{r}$ es el error residual del análisis de varianza (ANAVA) combinado (que incluye todas las pruebas regionales):

$$\sum_j \hat{\delta}_{ij}^2 = \left[\sum_j Y_{ij}^2 - Y_i^2 \right] - \frac{\left(\sum_j Y_{ij} I_j \right)^2}{\sum_j I_j^2} \quad [5]$$

Así, este modelo permite una forma de particionar la interacción genotipo-ambiente de cada cultivar en dos partes:

- desviación debida a la respuesta de la variedad al cambio en índice ambiental (suma de cuadrados de la regresión) y;

- desviaciones no explicadas por el modelo; las cuales son una aproximación a la estabilidad fenotípica del cultivar.

El concepto de estabilidad descrito, permite clasificar los cultivares como lo propone Vallejo y Estrada (2002) y se presenta en la Tabla 1.

Tabla 1. Significado de los parámetros de estabilidad obtenidos por la metodología de Eberhart y Russell.

Coefficiente de Regresión	C.M. de la desviación de la ($\sum d_{ij}^2$) regresión	Significado
= 1	= 0	Variedad estable y predecible
= 1	> 0	Buena respuesta en todos los ambientes, pero poco predecible
< 1	= 0	Mejor respuesta en ambientes desfavorables y predecible.
< 1	> 0	Mejor respuesta en ambientes desfavorables, pero poco predecible.
> 1	= 0	Mejor respuesta en ambientes favorables y predecible.
> 1	> 0	Mejor respuesta en ambientes favorables, pero poco predecible.

Fuente: Vallejo y Estrada (2002)

Modelo de componentes de varianza de Shukla

Este propone la estimación de componentes de varianza dentro del efecto de la interacción genotipo ambiente, con un modelo lineal de la forma propuesta por Piepho (1999):

$$y_{ijk} = \mu + g_i + r_{jk} + u_j + (g\varepsilon)_{ij} + e_{ijk} \quad [6]$$

donde y_{ijk} ($i=1, \dots, I$ genotipos; $j=1, \dots, J$ ambientes; $k=1, \dots, K$ repeticiones) es la respuesta en la k -ésima repetición del i -ésimo genotipo en el j -ésimo ambiente, μ es la media general, g_i el efecto del i -ésimo genotipo, ε_j es el efecto del j -ésimo ambiente, $(g\varepsilon)_{ij}$ es el efecto de la interacción del genotipo i en el ambiente j , y e_{ijk} es el término del error experimental asociado a y_{ijk} (Shukla 1972).

Los efectos aleatorios r_{jk} , ε_j , $(g\varepsilon)_{ij}$ y e_{ijk} son asumidos independientemente distribuidos con media cero y varianzas $\sigma_{r_{jk}}$, σ_{ε_j} , $\sigma_{(g\varepsilon)_{ij}}$ (estabilidad de varianzas) y $\sigma_{e_{ijk}}$ respectivamente (Piepho 1999). Trabajando con los promedios, el modelo $y_{ijk} = \mu + g_i + \varepsilon_j + (g\varepsilon)_{ij} + e_{ijk}$ se define como (Shukla 1972):

$$\bar{y}_{ij} = \mu + g_i + \varepsilon_j + (g\varepsilon)_{ij} + \bar{e}_{ij} \quad [7]$$

Se asume que $E(\bar{e}_{i,j}) = 0$; $V(\bar{e}_{i,j}) = \sigma_0^2$, para todo i y j , donde \bar{e}_{ij} es la media de los e_{ijk} y σ_0^2 es la varianza del error dentro de los ambientes. Además también se asume que los efectos ambientales son aleatorios con media poblacional cero y varianza σ_ε^2 . Se selecciona una muestra aleatoria de S ambientes en una infinita

población de ambientes. σ_0^2 es estimada por $\hat{\sigma}_0^2$ (Shukla 1972), donde

$$\hat{\sigma}_0^2 = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij})^2 / str(r-1) \quad [8]$$

Con $st(r-1)$ grados de libertad. En adelante se trabaja con el promedio de los \bar{y}_{ij} (la media de los genotipos en cada ambiente).

La caracterización de los genotipos con base en los coeficientes de regresión puede no ser muy efectiva cuando solo una pequeña fracción de esta suma de cuadrados (GxA) se puede atribuir a la heterogeneidad entre las regresiones. Por esto, resulta de interés la partición de ésta en t componentes, cada uno atribuible a cada genotipo. Haciendo $(g\mathcal{E})_{ij} + \bar{e}_{ij} = v_{ij}$ y asumiendo que $E(v_{ij}) = 0$; $V(v_{ij}) = \sigma_i^2$; $E(v_{ij}, v_{i'j'}) = 0$ para todo $i \neq i'$ o $j \neq j'$; $V(g\mathcal{E}_{ij}) = \sigma_i^2$; también

$$E(g_{ij}, \bar{e}_{ij}) = 0; \quad i = 1, \dots, t. \quad [9]$$

Luego,

$$\sigma_i^2 = \sigma_i^2 + \sigma_0^2 \quad [10]$$

En esta expresión σ_i^2 puede ser tomado como la suma de dos componentes, varianza ambiental (σ_0^2) y varianza ambiental del genotipo i (después de hacer la corrección para el efecto aditivo común del ambiente \mathcal{E}_j), y se denomina estabilidad de varianza ("*stability variance*") del i -ésimo genotipo (Shukla 1972). Se va denominar un

genotipo estable a aquel que tenga estabilidad de varianza igual a la varianza ambiental (σ_0^2) lo que significa que $\sigma_i^2 = 0$. Valores relativamente grandes de σ_i^2 pueden ser indicadores de más inestabilidad ambiental del genotipo.

Evitando entrar en detalles, se obtiene una estimación insesgada de σ_i^2 , denotada por $\hat{\sigma}_i^2$, como (Shukla 1972):

$$\hat{\sigma}_i^2 = K \left[t(t-1) \sum_j (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - \sum_i \sum_j (y_{ij} - \bar{y}_{i.} + \bar{y}_{..})^2 \right]$$

$$\hat{\sigma}_i^2 = K \left[t(t-1) \sum_j (u_{ij} - \bar{u}_{i.})^2 - \sum_i \sum_j (u_{ij} - \bar{u}_{i.})^2 \right] \quad [11]$$

Donde $K = \frac{1}{(s-1)(t-1)(t-2)}$, $u_{ij} = \bar{y}_{ij} - \bar{y}_{.j}$ y $\bar{u}_{i.} = \sum_j \frac{u_{ij}}{s}$. Éstos son obtenidos como combinaciones lineales de la suma de cuadrados de los residuales $(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$, de esta forma, son independientes de μ , d_i y de la varianza de \mathcal{E}_i . No es muy difícil verificar que éstos son estimadores insesgados de σ_i^2 (Shukla, 1972). Bajo el supuesto de tener una distribución simétrica para los σ_i^2 , Rao (1970) ha probado que en promedio, que tienen una varianza mínima, entre todos los posibles estimadores cuadráticos insesgados (MINQUE) de σ_i^2 . Tampoco es difícil ver que su promedio es el mismo que la media de la suma de cuadrados de la interacción genotipo ambiente (GxA). De esta forma, al multiplicar

cada $\hat{\sigma}_i^2$ por $(t-1)(s-1)/t$ se obtienen t componentes de GxA, cada uno correspondiente a cada genotipo (Shukla 1972). Estos componentes no son estadísticamente independientes; por ser éstos diferencias de dos sumas de cuadrados, pueden ser negativos, pero estimados negativos de varianzas no son poco comunes en problemas de componentes de varianzas.

La varianza de $\hat{\sigma}_i^2$ no es sólo función de σ_i^2 sino también de las varianzas de los otros genotipos σ_j^2 ($j \neq i$) tomados en el ensayo. Tales estimadores solo son disponibles cuando $t \geq 3$.

Resulta de interés probar si ciertos genotipos son estables o no. Johnson (1962) sugiere un criterio de prueba basado en la razón $\hat{\sigma}_i^2 / \hat{\sigma}_0^2$.

Es difícil derivar la distribución exacta de $\hat{\sigma}_i^2$, pero cuando t es grande, la varianza de $\hat{\sigma}_i^2$ se estima aproximadamente como lo propone Shukla (1972):

$$V(\hat{\sigma}_i^2) \cong 2\sigma_i^4 / (s-1) \quad [12]$$

La anterior expresión ayuda a obtener una distribución aproximada. Cuando t es grande, $(s-1)\hat{\sigma}_i^2 / \sigma_0^2$ será aproximadamente distribuida como una χ^2 con $(s-1)$ grados de libertad (Shukla 1972); así, bajo la hipótesis de que $\sigma_i^2 = 0$, F^* tendrá una distribución F aproximada con $(s-1)$ y $st(r-1)$ grados de libertad donde

$$F^* = \hat{\sigma}_i^2 / \hat{\sigma}_0^2 \quad [13]$$

Cuando $\hat{\sigma}_i^2$ es negativa o menor que $\hat{\sigma}_0^2$ entonces $\hat{\sigma}_i^2$ puede ser tomada como igual a cero como es usual.

La necesidad de esta aproximación³ es una de las razones por las cuales en este trabajo se recurre al estimador tipo Jackknife para calcular la estabilidad fenotípica.

La técnica de validación cruzada

Esta técnica se recomienda como una de las mejores formas de probar un modelo, pues es bien sabido que se introduce un sesgo al probar la validez de un modelo con los mismos datos con los que fueron estimados sus parámetros (Cooil et al 1987). La validación cruzada muchas veces es tomada como sinónimo de partición de los datos, pero según este autor, existen tres aproximaciones a la misma: i) Partición de los datos, la cual separa la estimación del modelo de la validación, ii) métodos de uso repetido de la muestra tales como el Jackknife, los cuales se concentran principalmente en la estimación de parámetros y, iii) los métodos que estos mismos autores llaman métodos simultáneos, en los cuales se estima y se valida el modelo.

Mosier (citado por Cooil 1987), define la validación cruzada como la prueba de efectividad en una parte de la

³ Al igual que la no independencia de los estimados de los $\beta_{k(i)}$ en el modelo de Eberhart y Russell.

muestra, tomada de forma independiente de la muestra total. Con el tiempo, el término ha llegado a significar el análisis de una muestra por división de la misma en submuestras.

Usualmente, el investigador divide la muestra en dos partes y a continuación, usa una parte llamada muestra de estimación para calcular los parámetros del modelo. La ecuación resultante es usada para predecir los valores de la variable dependiente del resto de las observaciones -las que se dejaron por fuera (Steckel & Vanhonaker 1993).

La aproximación a la validación cruzada en forma de subdivisión de la muestra puede ser rastreada hasta Larson (1931) mas aun no se plantean respuestas a preguntas como: i) ¿Qué reglas formales o heurísticas existen para decidir si el modelo estimado es estable? Y, ii) ¿Qué proporción de la muestra dedicar a la estimación sin ir contra el principio de suficiencia?

Un tratamiento más formal sobre la validación cruzada y en términos de espacios coordenados, sobre los que se estiman familias de modelos lineales se encuentra en Venter & Snyman (1995).

Jackknife

Esta técnica pueden ser usadas para estimar la precisión de coeficientes, sobre todo cuando no se conoce la distribución exacta para estimar la varianza de los parámetros estimados y/o la distribución muestral es

desconocida. Esta técnica -usada para responder a la pregunta "¿Qué tan preciso es mi estimador?"- puede ser usada para estimar sesgo o error estándar, mas no para estimar intervalos de confianza (Efron 1982); Manly (1996), establece que la técnica falla cuando el estimador no se distribuye normalmente; sin embargo, puede mejorar su desempeño cuando se utilizan transformaciones que lo normalicen; este autor sugiere trabajar con la transformación logaritmo natural. Dixon (1993) añade que intervalos de confianza basados en error estándar estimado con Jackknife asumen que el estadístico tiene una distribución normal y que si esta distribución es asimétrica o tiene colas pesadas, los intervalos son imprecisos.

La idea básica en el proceso Jackknife es que el sesgo y el error estándar de un estadístico pueden ser estimados por el recálculo del estadístico en un subgrupo de los datos (Dixon 1993), lo que lo convierte en una forma de validación cruzada. Algunos resultados teóricos muestran que en muestras extremadamente grandes las técnicas de Bootstrap y Jackknife arrojan resultados similares (Efron 1982 citado por Dixon 1993).

La palabra "jackknife" describe un gran número de herramientas que pueden ser usadas en muchas tareas, si bien, raramente es ideal para alguna de ellas (Manly 1996). La palabra "jackknife" fue propuesta por Tukey (1958) para usar en estadística al describir una aproximación general para pruebas de hipótesis y el cálculo de intervalos de confianza en situaciones en las

cuales otros métodos -mejores métodos- no pueden ser fácilmente usados.

Una forma de justificar el uso de los estimadores "Jackknife" es pensar en términos de lo que se hace con la estimación de la media, pero desde un punto de vista inusual. Suponga que se tiene una muestra aleatoria de n valores X_1, X_2, \dots, X_n y se estima la media de la población

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Suponga que este promedio es calculado otra vez, pero con la j -ésima observación perdida arrojando:

$$\bar{X}_{-j} = \frac{(\sum_{i=1}^n x_i - x_j)}{n-1} \quad [14]$$

Las dos últimas ecuaciones pueden ser resueltas para X_j resultando:

$$X_j = n\bar{X} - (n-1)\bar{X}_{-j}$$

De esta forma, es posible determinar el valor de X_j de la media general y la media excluida la j -ésima observación.

En general, suponga que un parámetro θ es estimado por alguna función de n valores muestrales, este estimado puede ser denotado por $\hat{E}(X_1, X_2, \dots, X_n)$, y escrito brevemente por \hat{E} . Con la exclusión de X_j , el estimado

parcial de θ es \hat{E}_{-j} . Por analogía con la ecuación [14], existe un conjunto de "pseudo valores"

$$\hat{E}_{-j}^* = n\hat{E} - (n-1)\hat{E}_{-j} \quad [15]$$

Para $j = 1, 2, \dots, n$. Estos pseudo valores juegan el mismo papel que los X_j en la estimación de la media y el promedio de los pseudo valores.

$$\hat{E}^* = \frac{\sum_{j=1}^n \hat{E}_j^*}{n} \quad [16]$$

Es el estimador jackknife de θ (Manly 1996). Si se tratan los pseudo valores como una muestra aleatoria de estimaciones independientes, entonces se sugiere que la varianza de este estimador jackknife puede ser estimada por S^2/n , donde S^2 es la varianza muestral de los pseudo valores estimados. Avanzando un poco más, se sugiere que un intervalo al $(1-\alpha)100\%$ para θ está dado por

$$\hat{E}^* \pm t_{\frac{\alpha}{2}, (n-1)} S / \sqrt{n}, \text{ donde } t_{\frac{\alpha}{2}, (n-1)} \text{ es el valor que excede la}$$

probabilidad de $\alpha/2$ en la distribución de t con $n-1$ grados de libertad (Manly 1996).

Se puede decir que se transforma un problema de estimación en la tarea de hallar un promedio muestral. Todo lo que se necesita es que el estimador del parámetro usual de interés sea una función de los valores de la muestra. En particular, los pseudo valores posiblemente están correlacionados y así la varianza estimada del estimador jackknife presenta alguna forma de sesgo y particularmente, predecir si es

el caso o no, es teóricamente difícil. Usualmente la varianza del estimador jackknife necesita ser justificada por estudios numéricos de rango antes que pueda ser confiable.

Antes de Tukey, Quenouille (1956) había señalado que el reemplazar un estimador por su versión "Jackknife" tenía el efecto de remover sesgo de orden $1/n$. Para verlo, suponga que el valor esperado (la media) del estimador \hat{E} de θ , basado en toda la muestra, sea $\theta(1+A/n)$. Entonces el valor esperado del estimador parcial \hat{E}_{-j} , con la j -ésima removida, es $\theta\{1+A/(n-1)\}$. Se sigue que el valor esperado del j -ésimo pseudo valor como es definido por la ecuación [18] es (Manly 1996):

$$n[\theta(1 + A/n)] - (n-1)[\theta\{1 + A/(n-1)\}] = 0,$$
 con el factor de sesgo simplificado. Así, para cada valor de j y consecuentemente para cada valor esperado de \hat{E}^* , el promedio de los pseudo valores es también θ .

Esta cualidad de los estimadores jackknife se pretende utilizar para los estimadores de estabilidad genotípica de Shukla y de Eberhart & Russell. En el anexo 1 se encuentra el código para el programa SAS^(R) con el que se realizaron los análisis completos y el proceso de obtener la estimación jackknife de las medidas de estabilidad mencionadas.

Metodología

Se utilizarán datos provenientes de 20 pruebas multiambientales, las cuales fueron establecidas bajo un diseño en bloques completos al azar con cuatro repeticiones para nueve líneas elites de mejoramiento, en cada localidad o ambiente.

Inicialmente se realizó un análisis combinado de las pruebas regionales según el siguiente modelo:

$$Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + ZU + e \quad [17]$$

Donde Y es el vector de observaciones ($n \times 1$), X_1 y β_1 son, respectivamente, la matriz de diseño ($n \times a$) y el vector ($a \times 1$) de efectos fijos de los ambientes ($j=1, \dots, a$); X_2 y β_2 son, respectivamente, la matriz de diseño ($n \times g$) y el vector ($g \times 1$) de efectos fijos de los genotipos ($i=1, \dots, g$); X_3 y β_3 son, respectivamente, la matriz de diseño ($n \times ag$) y el vector ($ag \times 1$) de efectos fijos de la interacción genotipo por ambiente, Z y U son, respectivamente, la matriz de incidencia ($n \times ab$) y el vector ($ab \times 1$) de efectos aleatorios de los bloques ($r=1, \dots, b$) dentro de los genotipos, y e es el vector ($n \times 1$) asociado al error. Se supone que $U \sim N(0, \sigma_{ab}^2)$ y $e \sim N(0, \sigma^2)$.

Para la implementación de las metodologías de Eberhart y Russell y varianza de Shukla se utilizó la varianza σ^2 estimada del modelo anterior. Para esto fue implementado en un módulo en el programa SAS^(R) que

realiza la rutina RDMS (anexo 1). Módulos adicionales permitieron la comparación por conteo de las veces en que el estadístico de Shukla resultó mejor que el de Eberhart & Russell.

Un MACRO elimina a cada combinación genotipo por ambiente, alguna de sus cuatro repeticiones; Un billón (1099511627776) de posibles subconjuntos puede ser tomado para estimar de nuevo la estabilidad. Los posibles subconjuntos diferentes que se pueden tomar son muestreados de forma aleatoria, y cuidando de no repetir una combinación de exclusiones; ésta forma de muestreo hace del procedimiento una forma de estimación tipo Jackknife. Para efectos de garantizar la estimación, siempre con una combinación diferente de repeticiones, le fue asignado a cada una de éstas una identificación numérica.

En el primer módulo de ese programa se escriben las operaciones de lectura y ordenación de la base de datos, se toma una muestra de n-1 elementos (el programa extrae subconjuntos siempre diferentes de este tamaño) y se llaman las rutinas almacenadas en el segundo módulo.

En el segundo módulo, se calculan los estadísticos (E&R y Shukla) y se estiman los valores esperados del rendimiento para cada genotipo y con éstos, la diferencia respecto de los valores estimados con todos los datos. Tanto los parámetros como la diferencia respecto de la información completa son almacenados en archivos separados; el programa retorna estos valores al módulo principal (Teste.sas).

En este punto se repite el proceso hasta el número especificado de muestreos.

El proceso de validación cruzada permite obtener la diferencia entre el valor observado y el estimado (independientemente). En el conjunto de tales diferencias, cada una se elevó al cuadrado y se acumuló en el estadístico RDMSP para cada método.

$$RDMSP_{Shukla} = \frac{\sum_{i=1}^{25000} \sqrt{\sum_{j=1}^{180} (\hat{Y}_{Shukla} - Y_{obs})^2}}{25000} \quad [18]$$

$$RDMSP_{Eberhart\&Russell} = \frac{\sum_{i=1}^{25000} \sqrt{\sum_{j=1}^{180} (\hat{Y}_{Eberhart\&Russell} - Y_{obs})^2}}{25000} \quad [19]$$

Con el conjunto de parámetros estimados se calcularon los percentiles 2.5 y 97.5 que permitieron hallar intervalos de confianza tipo Jackknife para los parámetros de estabilidad vía Shukla y Eberhart & Russell.

Posteriormente, se realizó la comparación de la eficiencia relativa de los dos métodos mediante los desvíos al cuadrado del estimado con la base de datos completa usando la suma de distancias descrita por [18] versus la descrita por [19].

También, fueron comparadas las diferencias entre la estimación puntual con los datos completos y la

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{ia} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_j \\ \vdots \\ A_a \end{bmatrix}, \quad e_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ij} \\ \vdots \\ e_{ia} \end{bmatrix}$$

donde y_i es el vector de medias de cada ambiente para el i -ésimo genotipo, \mathbf{A} es el vector de índices ambientales para el ambiente j -ésimo, $\mathbf{1}_a$ es un vector columna de unos, de tamaño "a", β_{oi} (el intercepto para cada genotipo) es el promedio de rendimiento para el i -ésimo genotipo, β_{li} (el coeficiente de regresión) es el índice de adaptabilidad asociado al genotipo i , y e_i es el vector de efectos del error para el i -ésimo genotipo. Se asume $e_i \sim N(0, \sigma_i^2)$.

Modelo de Shukla

Este método también utiliza el modelo de medias de genotipos por ambiente de la siguiente forma:

$$Y = X_1\beta_1 + X_2\beta_2 + e \quad [21]$$

Donde Y es el vector ($ag \times 1$) de promedios de los genotipos en cada ambiente, X_1 y β_1 son, respectivamente, la matriz de diseño ($ag \times a$) y el vector ($a \times 1$) de efectos fijos de los ambientes, X_2 y β_2 son, respectivamente, la matriz de diseño ($ag \times g$) y el vector

de efectos fijos de los genotipos; $e = [e_1, \dots, e_i, \dots, e_g]$ de efectos aleatorios de la interacción genotipo por ambiente. Se asume que $e_i \sim N(0, \sigma_i^2)$.

Resultados y discusión

Se realizaron 25000 muestreos excluyendo alguna combinación diferente de repeticiones eliminadas en cada celda de la tabla Genotipo Ambiente. El tiempo real de computo se estima en 28 horas en una maquina Pentium 4, con una giga de memoria RAM. Un resumen de la estimación se halla en la Tabla 2.

Tabla 2. Estadísticos básicos para el valor del RDMSP al realizar 25000 iteraciones del proceso de validación cruzada sobre los datos de trigo.

Método	\bar{X}	S	Máximo	Mínimo	L. Inf.	L. Sup.
Shukla	13285.04	648.59	15632	10627.66	13235.12	13334.96
Eberhart & Russell	12997.56	660.76	15341.64	9998.89	12947.18	13047.94

Con intervalos relativamente equivalentes en amplitud, el estadístico RDMSP para el método de Eberhart & Russell es más efectivo al estimar el futuro valor de los rendimientos de un genotipo en un determinado ambiente. La dispersión en ambos estadísticos es aproximadamente la misma, pero los intervalos no se traslapan. En promedio, los métodos difieren en 1000 unidades como se observa en las figuras 3 y 4, siendo el método de Eberhart & Russell el que mejor estimaciones produce.

Tabla 3. Pendiente de la recta estimada por el estadístico de Eberhart & Russell, con base en 25000 remuestreos sin reemplazo.

Genotipo	Percentil 2.5	Percentil 97.5	Estimación puntual
G1	0.9080	1.0551	0.98139452
G2	0.8878	1.0493	0.96505717
G3	0.8598	1.0474	0.95400058
G4	1.2182	1.4279	1.31430365
G5	1.1402	1.3125	1.22808766
G6	0.9441	1.1684	1.0574938
G7	0.7760	0.9378	0.85756455
G8	0.4494	0.5992	0.52397566
G9	1.0158	1.2170	1.11812242

Según los valores de estabilidad fenotípica, β_i , los genotipos 4, 5 y 9 presentan mejor respuesta en ambientes favorables; los genotipos 7 y 8 tienen mejor respuesta en ambientes desfavorables. Los demás genotipos son clasificados como estables (tabla 3 y Figura 1).

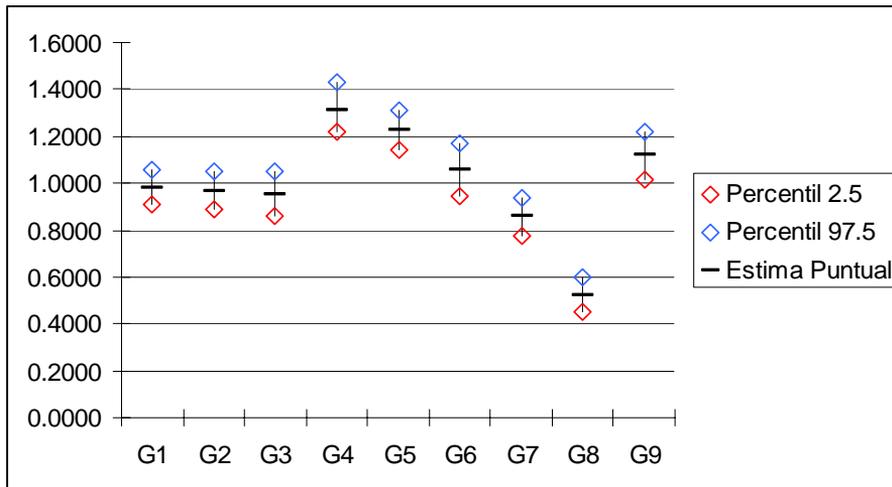


Figura 1. Intervalos de confianza al 95% para la estabilidad genotípica de nueve genotipos de trigo ensayados en 20

localidades. Los genotipos 4, 5 y 9 presentan mejor respuesta en ambientes favorables.

Otra forma de verificar el desempeño de los métodos de Shukla y Eberhart & Russell puede ensayarse sustrayendo de la base de datos otros subconjuntos de observaciones (una manera de hacer esto es cambiando la semilla "seed()" en el código) -extrayendo más de una repetición por celda para conocer el comportamiento en ensayos desbalanceados, situación que este trabajo no considera.

Las figuras 2 y 3 muestran la distribución de los estadísticos de estabilidad calculados por el método de Eberhart & Russell, y la varianza de estabilidad fenotípica de Shukla.

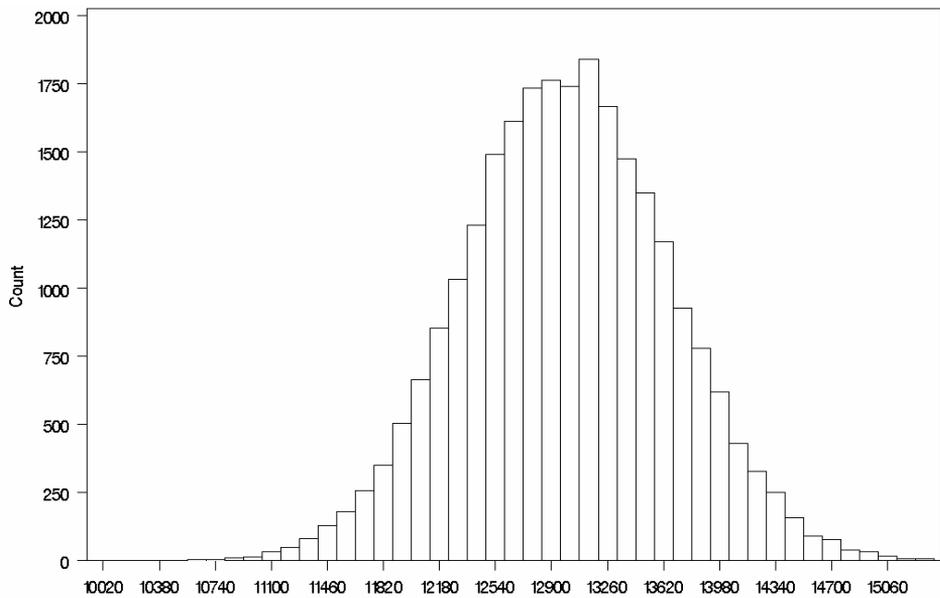


Figura 2. RDMSP para el estadístico de Eberhart & Russell

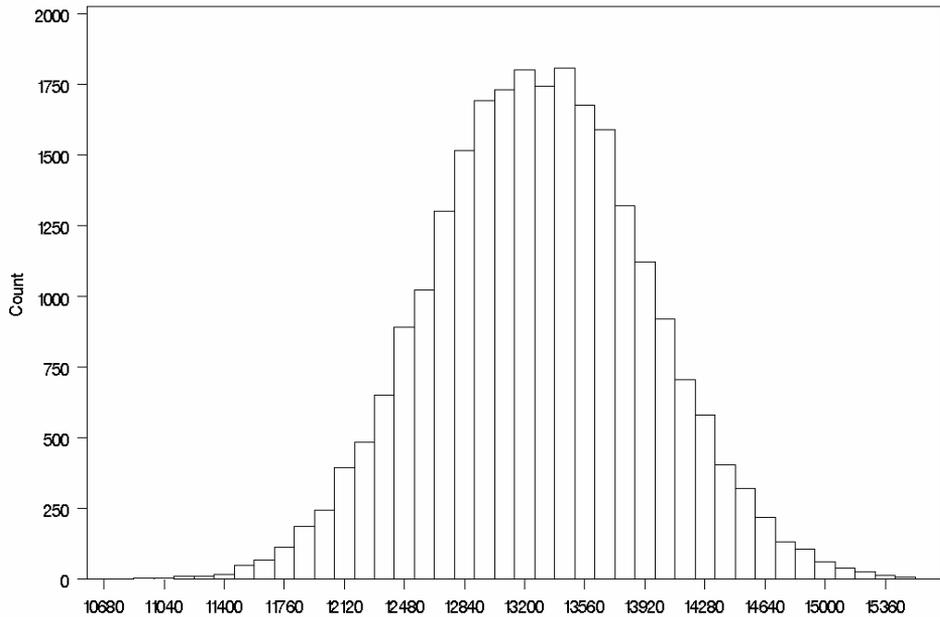


Figura 3. RDMSP para el estadístico de Shukla

Utilizando el esquema de validación cruzada se estimaron por el método de jackknife los intervalos de confianza para los parámetros de estabilidad fenotípica de Eberhart & Russell y de Shukla. La tabla 4 muestra la estimación puntual con los datos completos de la varianza en la estimación del parámetro de estabilidad de Eberhart & Russell.

Tabla 4. Sigma-i para cada genotipo estimado por el estadístico de Eberhart & Russell con base en 25000 remuestrs sin reemplazo.

Genotipo	Percentil 2.5	Percentil 97.5	Sigma i
G1	181754.34	499400.26	336926.285
G2	0.00	178168.56	87267.1056
G3	0.00	90093.62	0
G4	132005.82	456532.32	289074.063
G5	18416.44	250068.36	134316.44
G6	0.00	162347.16	61646.4955
G7	21292.92	237488.94	131417.849
G8	103100.83	398915.05	251205.412
G9	29533.53	350769.06	183460.79

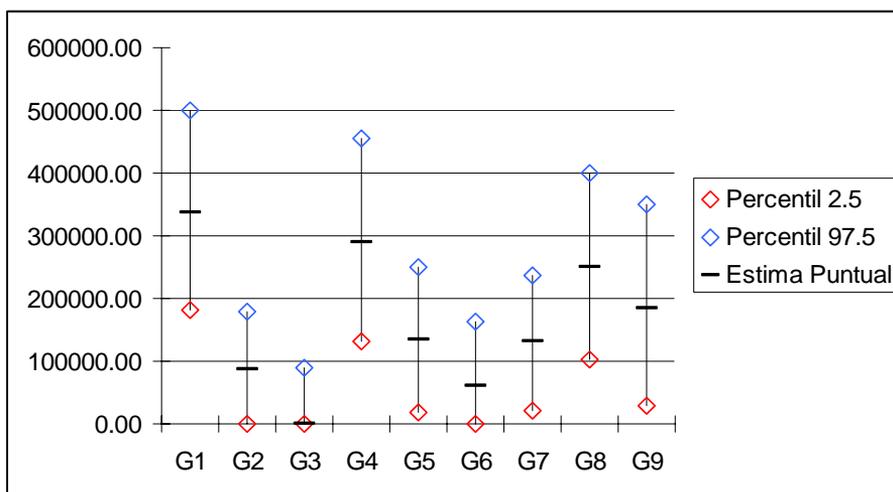


Figura 4. Intervalos de confianza al 95% para Sigma en cada genotipo en la estimación de Eberhart y Russell, con base en 25000 remuestros sin reemplazo. Los genotipos 2, 3 y 6, cuyos intervalos contienen el cero, se clasifican como estables.

Como se observa, los intervalos son bastante amplios, lo cual concuerda con lo reportado por Efron (1982). Sin embargo, son una buena alternativa para su utilización, en este caso, porque éstos dan una buena respuesta al problema de las aproximaciones presentadas originalmente por los autores de estos métodos.

Tabla 5. Estimación de intervalos de confianza al 95% para la varianza de Shukla, con base en 25000 remuestros sin reemplazo.

Genotipo	Percentil 2.5	Percentil 97.5	Estimación puntual
G1	1279237.36	2357446.67	2143751.64
G2	546908.12	1239125.99	933876.337
G3	243765.77	956868.99	501068.847
G4	1539437.13	2964306.33	2643059.81
G5	927492.85	1764221.68	1541111.52
G6	563817.92	1170913.87	824558.204
G7	761271.57	1588103.49	1290854.72
G8	1992492.30	3417850.65	3409551.99
G9	796806.99	1903565.97	1497281.76

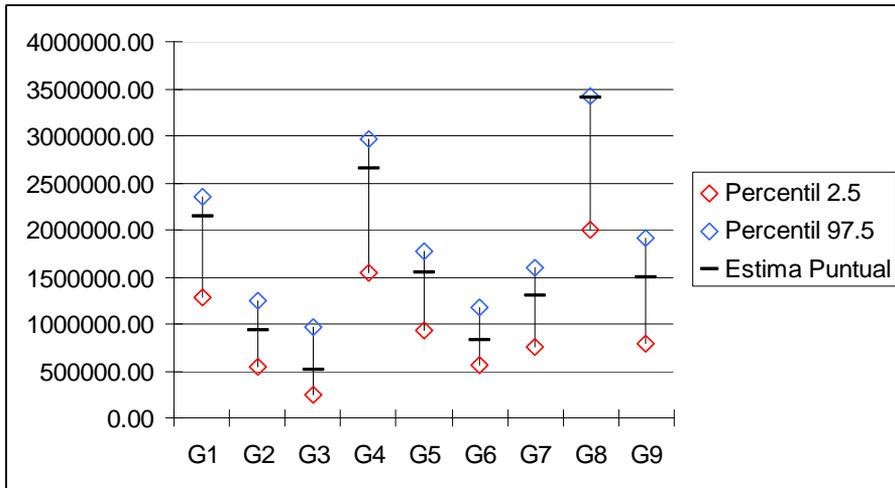


Figura 5. Intervalos de confianza al 95% para la varianza de Shukla, con base en 25000 remuestreos sin reemplazo.

REFERENCIAS

- Cornelius, P.L., and J. Crossa.** (1995). Shrinkage estimators of multiplicative models for crops cultivar trials. Technical report No. 352, University of Kentucky Department of Statistics, Lexington, Kentucky.
- Cornelius, P.L., J. Crossa and M.S. Seyedsadr.** (1994). Tests and estimators of multiplicative models for variety trials. Proceedings of the 1993 Kansas State University conference on applied statistics in agriculture. Manhattan, Kansas.
- Cooil, B., Winer, R. and Rados, D.** (1987) Cross-Validation for prediction. *Journal of Marketing Research* **XXIV**, 271-279.
- Cruz, C. D., Regazzi, J.** 2001. Modelos biométricos aplicados ao melhoramento genético. 2ª Ed.rev. Editora UFV. 390p.
- Dixon, P.M.** (1993) The bootstrap and the jackknife: Describing the precision of ecological indices, **In:** Design and analysis of ecological experiments. Editors: Samuel Scheiner and Jessica Gurevitch. Ed. Chapman & Hall. 445p.
- Eberhart, S.A. and Russel, W.A.** (1966). Stability parameters for comparing varieties. *Crop Science* **6**, 36-40.
- Efron, B.** (1982) The Jackknife, the bootstrap and the others resampling plans. Society of industrial and applied mathematics. CBMS-NSF Monograph 38.
- Finlay, K.W. and Wilkinson. G.N.** (1963). The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research* **14**, 742-754.
- Jinks, J.L. and Perkins, j.M.** (1968) Environmental and genotype environmental components of variability. III Multiple lines and crosses. *Heredity* **23**, 239-256.
- Johnson, N.L.** (1962). Some notes on the investigation of heterogeneity in interaction. *Trabajos de Estadística*, **XIII**, 183-199.
- Larson, S.C.** (1931) The shrinkage of the coefficients of multiple correlation. *Journal of Educational psychology* **22**, 45-55.
- Mandel, j.** (1961). Non-additivity in two way analysis of variance. *J.Am.Stat.Assoc.*, **56**, 878-888.
- Manly, F.J.** (1996) Randomization, Bootstrap and Montecarlo Methods in biology. 2º Ed. University of Otago, New Zealand. Chapman & Hall. London-Madras. 455p.
- Mosier, C.I.** (1951) Problems and design of cross-validation. *Educational and psychological measurement* **11**, 5-11.
- Piepho, H.P.** (1999) Stability analysis using the SAS System. *Agronomy Journal* **91**, January-February, 154-160.
- Steckel, J.H. and Vanhonaker, W. R.** (1993) Cross-validating regression models in marketing research. *Marketing Science* **12**(4), 415-427.
- Quenuoville, M. H.** (1956) Notes on bias in estimation. *Biometrika*, **43**, 353-60.

- Venter, J.H. and Snyman, J.L.J.** (1995) A note on the generalised cross-validation criterion in linear model selection. *Biometrika* **82**, 215-219.
- Rao, C.R.** (1970) Estimation of heteroscedastic variances in linear models. *Juornal of American statistics Assoc.*, **65**, 161-172.
- Shukla, G.K.** (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity* **29**, 237-245.
- Tai, G.C.C.** (1971) Genotype stability analysis and his application to potato regional trials, *Crop Science*, **11**, 184- 190.
- Tukey, J.W.** (1949). On degree of freedom of no aditivity. *Biometrics*, **5**, 232-242.
- Vallejo, F.A. y Estrada, E.I.** (2002). Mejoramiento genético de plantas. Universidad Nacional de Colombia (sede Palmira) 401p.
- Yates, F. and Cochran, W.G.** (1938) The analisis of groups of experiments. *J. Agric. Sci.*, **28**, 556-580.
- Freedman, G.H. and Perkins, J.M.** (1971) Enviromental and genotype-enviromental components of variability. *Heredity* **29**, 15-23.
- SAS OnLine Doc.** (1987). The IML Procedure. The SAS Institute.

ANEXO

TESTE . SAS :

```
dm "log;clear;output;clear";
libname jairo "C:\Jairo";
%include "C:\Jairo\pgm2.sas";
data jairo.trial;
  infile "C:\Jairo\EVT16B_84.dat";
  input env $ gen$ rep $ yraw;
run;
*****;
*Nota el archivo debe estar ordenado por gen, env y rep;
*****;
proc sort data=jairo.trial;
  by gen env rep;
run;
%macro remover(in,out,criterio);
  p=loc(&criterio);
  y=remove(t(&in),p);
  &out=y`;
%mend remover;
proc iml;
load module=_all_;
use jairo.trial;
read all;
start split;
  aleat=j(180,1,0);
  do i=1 to 180;
    aleat[i,]=int(1+4*ranuni(12345));
  end;
  tmp=j(4,1,1);
  aleat2=aleat@tmp;
  tmp=num(rep);
  h=(tmp=aleat2);
  h2=(tmp^=aleat2);
  %remover(yraw,yraw1,h);
  %remover(gen,gen1,h);
  %remover(env,env1,h);
  rep1=repeat({1,2,3},180,1);
  %remover(yraw,yraw2,h2);
  %remover(gen,gen2,h2);
  %remover(env,env2,h2);
  rep2=j(180,1,1);
  cromosoma=num(rowcatc(char(aleat)`));
  free tmp h h2 aleat2 aleat;
finish;
start crossvalidation;
  run
error(env1,gen1,rep1,yraw1,gen_m1,env_m1,total1,average1,mse1,dfc1,r1);
  run shukla(mse1, dfc1, r1, gen_m1, env_m1,
total1,out1_shukla1,out2_shukla1,y_sh1);
  run er(mse1, dfc1, r1, gen_m1, env_m1,
average1,out1_ER1,out2_ER1,y_er1);
```

```

run
error(env2,gen2,rep2,yraw2,gen_m2,env_m2,total2,average2,mse2,dfе2,r2);
run shukla(mse2, dfе2, r2, gen_m2, env_m2,
total2,out1_shukla2,out2_shukla2,y_sh2);
run er(mse2, dfе2, r2, gen_m2, env_m2,
average2,out1_ER2,out2_ER2,y_er2);
finish;
do k=1 to 10000; *< Colocar cuantas veces : 1000 iterac tardan 20 min; do
until(control=1); *Esto es para evitar que se repita dos veces la misma
secuencia;

run split;
v=cromosoma;
verif=0;
if k>1 then verif=verif//v;
if (verif=v) then control=0;else control=1;

end;
run crossvalidation; *Aqui físicamente es correr la validación
cruzada;
*Aqui de preparan los archivos de salida;
s1=k||out1_shukla1||out2_shukla2;
shukla=shukla//s1;
s2=k||y_sh1`||yraw2`;
s_pred=s_pred//s2;
s3=k||out1_er1||out2_er2;
er=er//s3;
s4=k||y_er1`||yraw2`;
er_pred=er_pred//s4;

end;

create shukla_val from shukla;
APPEND from shukla;
close;
create shukla_pred from s_pred;
APPEND from s_pred;
close;
create er_val from er;
APPEND from er;
close;
create er_pred from er_pred;
APPEND from er_pred;
close;

quit;
%macro rdmsp(method);
data RDMSp_pred_&method;
set &method._pred;
array cross (180) col2-col181;
array val (180) col182-col361;
array difference (180) dif1-dif180;
do i=1 to 180;
difference{i}=(cross{i}-val{i})**2;
end;
RDMSp=sqrt(sum(of dif1-dif180));
run;
proc means data=RDMSp_pred_&method;

```

```

        var RDMSP;
        title "&Method";
run;
quit;
title;
%mend rdmsp;
%rdmsp(Shukla);
%rdmsp(ER);

%macro limits(method,clase,begin,fin,correction);
proc delete data=limits_&method._&clase; run;
quit;
%do i=&begin %to &fin;
    data work;
        informat name $12.;
        set &method._val;
        y=col&i;
        %let j=%eval(&i-&correction-1);
        name="G&j";
    run;
proc univariate data=work noprint;
    var y;
    output out=limits pctlpre=P_ pctlpts=2.5 , 97.5;
    by name;
run;
proc append base=limits_&method._&clase data=limits;
run;
quit;
%end;
%mend limits;
%limits(Shukla,Variance,2,10,0);
%limits(ER,Average,2,10,0);
%limits(ER,Beta,11,19,9);
%limits(ER,Sigmai,38,46,36);

```

PGM. SAS:

```

dm "log;clear;output;clear";
proc iml;
start error(a,b,c,d,g,f,h,l,m,o,r);
    free f g h l m o r;
    *****Input*****
    a=>environment
    b=>genotype
    c=>block
    d=>response
    *****Output*****
    g=>genotype vector of the mean
    f=>enviroment vector of the mean
    h=>Vector with total of genotype by environment
    l=>Vector with mean of genotype by environment
    m=>Mean Square Error
    o=>Degree of Freedom Error
    r=>Number of block

```

```

*****
*A continuación se obtiene el cuadrado medio del error de un análisis de
varianza combinado *
*****;
n=nrow(d);
mu=j(n,1,1);
d_a=design(a);
d_c=design(c);
d_c_a=hdir(d_c,d_a);
d_b=design(b);
d_b_a=hdir(d_b,d_a);
X=mu||d_a||d_c_a||d_b||d_b_a;

bhat=ginv(X`*X)*X`*d;
ypred=X*bhat;
e=ypred-d;
sse=e`*e;
o=ncol(d_a)*(ncol(d_c)-1)*(ncol(d_b)-1);
if o=0 then m=.; else m=sse/o;
r=ncol(d_c);
*****
* A continuación se obtienen los h por genotipo y ambiente *
*****;
h=t(ypred`*d_b_a);
l=h/r;
*****;
*A continuación se obtienen los nombres para el archivo de medias, lo
importante es que el archivo original este ordena como se indica al principio de
este programa y que todos los genotipos tengan igual número de repeticiones, es
decir, no acepta datos faltantes;
*****;
do i=1 to n by r;
    z=b[i];
    e=a[i];
    g=g/z;
    f=f/e;
end;
free mu X bhat n d_d_c d_c_a d_b d_b_a;
finish error;
*****
Aquí inicia Shukla **
*****;
start shukla(m, o, r, g, f, h, o1, o2,o3);
free o1 o2 o3;
*****Input*****
m=>Mean Square Error
o=>Degree of Freedom Error
r=>Number of block
g=>genotype vector of the mean
f=>enviroment vector of the mean
h=>Vector with total of genotype by environment
*****Output*****
o1=>Row vector with Shukla's variance , F test and p value

```

```

o2=>Row vector with Shukla's variance
o3=>Predicted values for genotype by environment;
ntotal=nrow(h);
uno=j(ntotal,1,1);
denv=design(f);
nenv=ncol(denv);
dgen=design(g);
ngen=ncol(dgen);
X=uno||denv||dgen;
bhat=ginv(t(X)*X)*t(X)*h;
med=X*bhat;
o3=med/r;
e=h-med;
ee=t(e)*e;
eg=dgen#e;
egeg=vecdiag(t(eg)*eg);
suma=ngen*(ngen-1)*egeg-ee;
factor=1/((nenv-1)*(ngen-1)*(ngen-2));
shukla=factor*suma/r;
Fi=shukla/m;
ProbFi=1-probf(Fi,ngen-1,o);
*A continuación se obtienen los nombres de los genotipos y se imprime los
resultados;
do i=1 to ntotal by nenv;
    z=g[i];
    genotipo=genotipo//z;
end;
*print genotipo Shukla Fi ProbFi;
o1=Shukla`||Fi`||ProbFi`;
o2=Shukla`;
free uno ntotal denv nenv dgen ngen X bhat med e ee eg egeg suma factor
Shukla Fi ProbFi z genotipo;
*Aqui se libera la memoria y se deja solo lo que se necesita;
finish shukla;
*run shukla(m, o, r, g, f, h, out_shukla);
*****Aqui termina Shukla*****;
*****
Aqui inicia Eberhart y Russell **
*****;
start ER(m, o, r, g, f, l, o1, o2, o3);
free o1 o2 o3;
*****Input*****
m=>Mean Square Error
o=>Degree of Freedom Error
r=>Number of block
g=>genotype vector of the mean
f=>enviroment vector of the mean
l=>Vector with mean of genotype by environment
*****Output*****
o1=>Row vector with Shukla's variance , F test and p value
o2=>Row vector with Shukla's variance
o3=>Predicted values for genotype by environment ;
ntotal=nrow(l); * N *;

```

```

uno=j(ntotal,1,1); * Matriz de unos *;
denv=design(f); * Matriz diseño para Ambientes *;
nenv=ncol(denv); * Numero de ambientes *;
dgen=design(g); * Matriz diseño para genotipos *;
ngen=ncol(dgen); * Numero de genotipos *;
sum_env=t(denv)*1; * Suma de 1 para cada ambiente *;
n_env=t(denv)*uno; * Numero de genotipos en cada ambiente *;
prom_env=sum_env/n_env; * Promedio de 1 para cada ambiente *;
prom_general=(t(uno)*1)/ntotal; * Promedio general *;
indice=prom_env-prom_general; * Indice de productividad *;
tmp=j(nenv,1,1); * unos para la matriz de diseño de regresion *;
X=tmp||indice;
XX=inv(t(X)*X); * inversa ( Equiv transpuesta Equiv ) *;
Y_mod=dgen#1; * Matrix de diseño regresiones por cada genotipo *;
XXinv_mod=i(ngen)@XX;
X_mod=i(ngen)@X;
XY=T(X_mod)*1;
bhat=XXinv_mod*XY; * Matriz de betas estimada *;
o3=X_mod*bhat;
i2=t(indice)*indice*r;
tmp=repeat(i(2),ngen);
tmp2=tmp#bhat;
tmp3=i(ngen)@j(1,2,1);
beta=tmp3*tmp2;
YY=t(Y_mod)*Y_mod;
tmp4=i(ngen)@j(2,1,1);
tmp5=tmp4#XY;
tmp6=tmp4#bhat;
tmp7=t(tmp6)*tmp5;
dif=vecdiag(r*(YY-tmp7));
oreg=nenv-2;
CMD=dif/oreg;
sigmai=(CMD-m)/r;
do i=1 to ngen;
    if sigmai[i]<0 then sigmai[i]=0;
end;
if m^=. then do;
    var_Beta1=m/i2;
    var_Beta0=m/(r*nenv);
    tbeta1=(beta[,2]-1)/root(var_Beta1);
    probbtbeta1=2*(1-probt(abs(tbeta1),o));
    tbeta0=(beta[,2]-0)/root(var_Beta0);
    probbtbeta0=2*(1-probt(abs(tbeta0),o));
    Fi=CMD/m;
    ProbFi=1-probf(Fi,oreg,o);
end;
else do;*Aqui es para cuando solo hay una repeticion;
    tbeta1=j(ngen,1,.);
    probbtbeta1=j(ngen,1,.);
    tbeta0=j(ngen,1,.);
    probbtbeta0=j(ngen,1,.);
    Fi=j(ngen,1,.);
    ProbFi=j(ngen,1,.);

```

```

        sigmai=j(ngen,1,.);
    end;
*A continuación se obtienen los nombres de los genotipos y se imprime los
resultados;
    do i=1 to ntotal by nenv;
        z=g[i];
        genotipo=genotipo//z;
    end;
    Promedio=Beta[,1];
    Adap=Beta[,2];
    Estab=sigmai;
    o1=promedio`||adap`||tbeta1`||probtbeta1`||estab`||Fi`||ProbFi`;
    o2=promedio`||adap`;
*free /m o r g f totales o1 o2; *Aqui se libera la memoria y se deja solo lo
que se necesita;
finish ER;
*****Aquí Eberhart y Russell *****;
store module=_all_;
quit;

```