UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE AGRONOMÍA

ANÁLISIS DE LA ESTRUCTURA DE COVARIANZA PARA LA INTERPRETACIÓN DE LA INTERACCIÓN GENOTIPO-AMBIENTE CON DATOS DESBALANCEADOS

por

Victor Manuel PRIETO HERNÁNDEZ

TESIS presentada como uno de los requisitos para obtener el título de *Magister* en Ciencias Agrarias opción Ciencias Vegetales

MONTEVIDEO
URUGUAY
Diciembre 2014

Tesis aprobada por el tribunal integrado por Ing. Agr. (Dr.) Jorge Franco (Presidente), Ing. Agr. (PhD.) Federico Condón, Ing. Agr. (PhD.) Ariel Castro y PhD. José Crossa, el 1ero de diciembre de 2014. Autor: Ing. Agr. Victor Prieto. Director: Dr. Juan Burgueño.

Dedico este trabajo a mi esposa, Gabriela.

AGRADECIMIENTOS

Al director de tesis, Dr. Juan Burgueño, quien me acompañó desde el origen del trabajo hasta su culminación, proceso para mi cargado de enseñanzas.

Al Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) en la persona del Dr. José Crossa, quien estuvo como director de la Unidad de Biometría de dicho centro cuando esta investigación tuvo su comienzo.

Al personal de la Unidad de Posgrado y Educación Permanente, los integrantes de la Comisión Académica de Posgrado y especialmente a la coordinación de la Opción de Ciencias Vegetales de la Maestría en Ciencias Agrarias.

A todos mis compañeros de trabajo (y amigos) del DBEC, Departamento de Biometría, Estadística y Computación de la Facultad de Agronomía. A la profesora Estela Priore un agradecimiento especial por su apoyo y seguimiento.

A nuestra facultad, institución que nos abrió las puertas para formarnos como profesionales, docentes y como investigadores.

Muchas gracias.

TABLA DE CONTENIDO

RESUMEN

En el marco de un programa clásico de evaluación de cultivares, la información que usualmente proviene de ensayos multiambiente es altamente desbalanceada por selección o por pérdida circunstancial de datos, es además frecuente la presencia de interacción genotipo por ambiente (GEI), y los supuestos de homogeneidad de varianzas de los términos de la GEI entre ambientes no se sustentan. El enfoque de modelos mixtos para el análisis estadístico de los datos con las características antes mencionadas provee un marco de análisis más flexible ya que: no depende de información completa, y es posible modelizar efectos fijos y efectos aleatorios, permitiendo a éstos últimos diferentes situaciones de homogeneidad/heterogeneidad a través de una estructura de varianzas-covarianzas más conveniente. De las diferentes estructuras existentes, la estructura de factores analíticos (FA) permite modelizar de forma más parsimoniosa la variación GEI. El presente trabajo tiene como objetivo el análisis de la estructura de covarianza e interpretación de la GEI en situación de incremento gradual de desbalance, generado a través de la simulación de diferentes grados de pérdida aleatoria de datos, bajo la hipótesis de que el desbalance generado tiene un efecto importante en el análisis de la GEI. Dicho efecto dependería del grado de pérdida en que se ven afectados nuestros datos, así como el tipo de información que se pierde (pérdida en parcelas, genotipos y/o sitios). Para llevar a cabo el trabajo se utilizaron datos de rendimiento en grano de diferentes ambientes y genotipos provenientes de ensayos de trigo semi-árido del CIMMYT. La comparación entre las estructuras de varianza se basó en la correlación del test de Mantel y la técnica de Procrustes. Surgen como resultados del trabajo una consistente pérdida de similaridad evidenciada en la caída de los valores de correlación entre la matriz completa y la que presenta desbalance, se pierde el patrón de respuesta de los diversos materiales y se manifiestan efectos de escala del ajuste Procrustes. Los resultados del presente trabajo permitirían a los investigadores tomar decisiones respecto a la confiabilidad que puedan presentar los resultados de sus análisis dependiendo de la pérdida de información que tengan.

Palabras claves: Interacción genotipo por ambiente, datos desbalanceados, modelos mixtos, factores analíticos, estructura de covarianza.

ANALYSIS OF THE COVARIANCE STRUCTURE FOR THE INTERPRETATION OF THE GENOTYPE-BY-ENVIRONMENT INTERACTION WITH UNBALANCED DATA.

SUMMARY

Under a classical cultivar evaluation program, the information usually comes from multi-environment trials that are highly unbalanced by selection or circumstantial data loss, often presenting genotype by environment interaction (GEI), and assumptions of homogeneity of variances of the GEI terms not supported. The mixed model approach for statistical analysis of the data with the characteristics mentioned above provides a more flexible framework because: does not depend on complete information, and is possible to model fixed and random effects, the latter allowing different homogeneity/ heterogeneity situations through a more convenient variance-covariance structure like the factor analytical (FA), which can model the GEI variation in a more parsimonious way. The aim of the present study was to analyze the covariance structure and interpretation of GEI through simulation of different random data loss schemes, under the hypothesis that this missing data has an effect on GEI analysis. The extent of the effect would depend on the level of missing data, and the type of information (plots, genotypes and / or sites) lost. A semi-arid CIMMYT wheat yield trial was used consisting of different genotypes and several international sites. The comparison between the variance structures was based on Mantel test correlation and Procrustes analysis. As a result of this work, a consistent loss of similarity between the complete matrix and with one presenting missing data was evidenced, differences in the response pattern of the materials under evaluation, and scale effects on the residual variability of Procrustes fitted model. The results of this study could allow researchers to make decisions about the reliability of their results that may arise depending on the information they have lost.

Keywords: genotype by environment interaction, unbalanced data, mixed models, factor analytic, covariance structure.

1. INTRODUCCIÓN

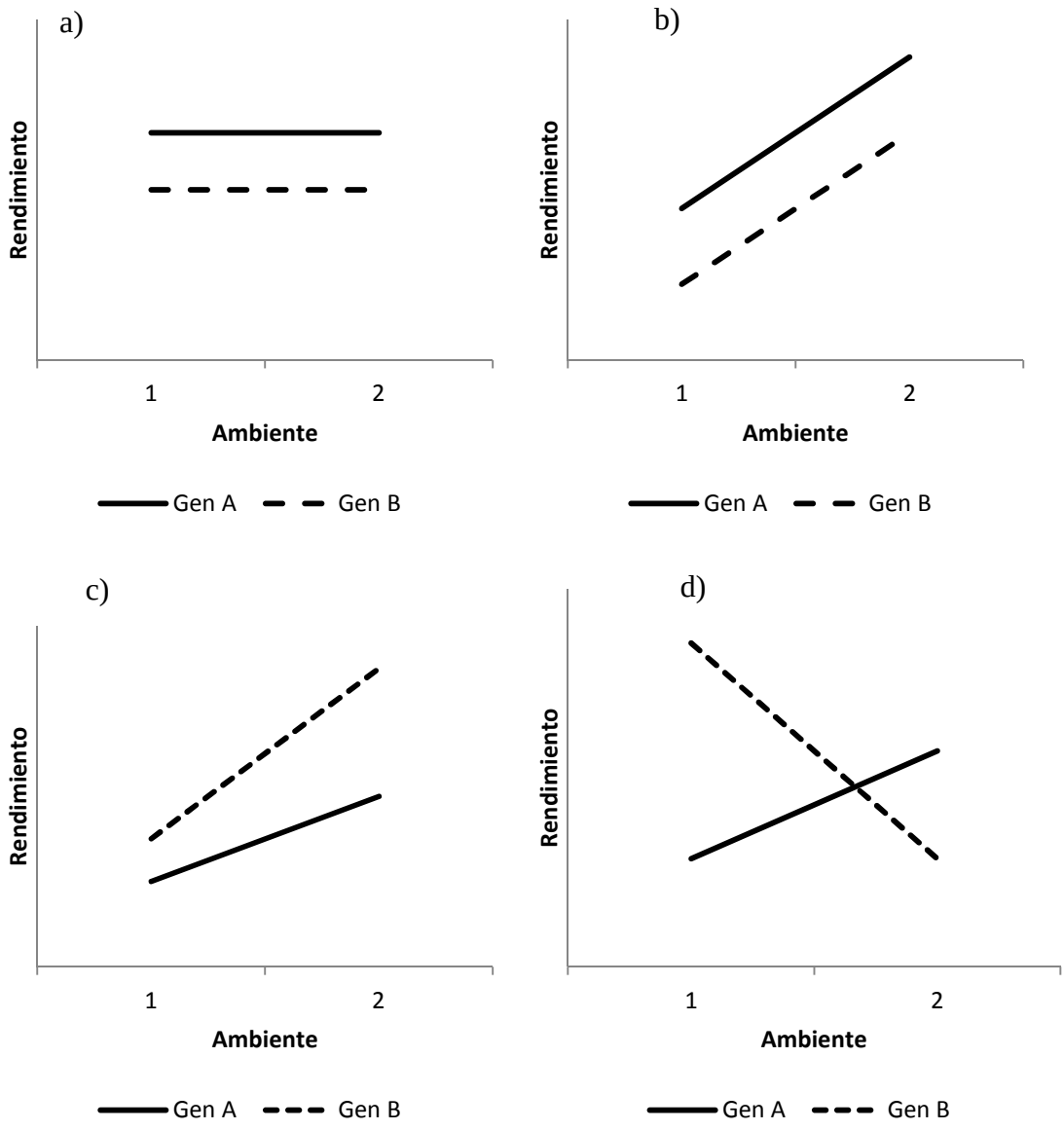
Uno de los principales propósitos en un programa de mejoramiento de cultivares es el estudio y cuantificación de la interacción genotipo-ambiente (*del inglés, Genotype by Environment Interaction* ó GEI). Una definición del término refiere a la respuesta diferencial de diferentes genotipos a través de diferentes ambientes (Kang, 2004). En las primeras revisiones sobre el tema (Hill 1975, Freeman 1973, Finlay y Wilkinson 1963) se citan trabajos de Fisher y Mackenzie (1923) y Yates y Cochran (1938) como las primeras referencias sobre el análisis de la GEI. Posteriormente hubo amplias revisiones (Xu 2010, Annicchiarico 2002; Hill *et al*. 1998, Kang y Gauch 1996, Crossa 1990) donde se profundiza en la interpretación y metodologías de análisis de la GEI.

1.1 INTERACCIÓN GENOTIPO-AMBIENTE

La expresión génica se sabe es dependiente de factores ambientales, y puede ser modificada, potenciada, y/o silenciada por los diferentes mecanismos reguladores de la célula en respuesta a fuerzas internas y externas. Los genotipos deben ajustarse a un estrés diferencial que se encuentra tanto dentro como entre ambientes. Los ambientes representan una muestra de años y sitios, e introducen grandes diferencias en la respuesta genotípica (Hanson, 1970). Cuando se evalúan genotipos vegetales, la variabilidad atribuible a la GEI es frecuentemente de una magnitud similar o mayor a la encontrada para la variabilidad genética.

Esta respuesta diferencial, medida en valores de caracteres observables (fenotipo) como rendimiento, altura de planta, resistencia a sequía, etc. puede ocasionar incluso el cambio de posición relativa o ranking de cultivares entre diferentes ambientes (Kang 1997, Kang y Gauch 1996) lo que complica la selección de los mejores materiales.

Esto puede ejemplificarse con la situación más simple posible: el caso de dos genotipos en dos ambientes:



En el diagrama vemos la situación donde (a) no hay respuesta del genotipo al cambio en el ambiente y (b) si la hay aunque ambos genotipos responden al cambio de ambiente en igual proporción manteniendo las diferencias que los separan, por ende no se manifiesta interacción. En (c) existe interacción ya que se observa la falta de paralelismo o simetría en la respuesta, pero los genotipos mantienen sus posiciones relativas entre sí. Es posible asignar un orden de jerarquía, por lo que la selección es clara. En la situación (d) donde también se manifiesta interacción, el orden de jerarquía de genotipos se invierte cuando pasamos del ambiente 1 al 2.

Mariotti (1994) define en su monografía (siguiendo a otros autores) como *interacción cualitativa* la que se da con cambio en el ranking de cultivares, mientras que la interacción sin ese cambio lo define como *interacción cuantitativa*. En este caso, existe respuesta diferencial (en valor absoluto) al cambio de ambiente, pero el orden jerárquico no cambia. Según el autor, los materiales fitotécnicos son mucho más difíciles de manejar cuando ocurren interacciones del tipo cualitativo, indicativas de que pueden ocurrir adaptaciones específicas de ciertos materiales a ciertas condiciones ambientales. En Ferreira *et al.* (2006) se ejemplifican también éstas situaciones de interacción donde en el caso de no existir un cambio en el ranking de genotipos, es llamada interacción simple. Si existiese cambio en el orden de los genotipos, es una interacción compleja, siendo ésta muy importante para el fitomejorador. La mayor parte de las situaciones reales es una combinación de casos de ausencia de interacción, interacciones simples e interacciones complejas.

Annicchiarico (2002) discute los efectos ambientales que reflejan el potencial ecológico de un sitio o de una determinada condición de manejo, que no son de importancia directa para el fitomejorador, mientras que el efecto del genotipo (es decir, diferencia media en el rendimiento entre genotipos) provee la única información relevante en ausencia de GEI (o cuando la misma es ignorada).

La creciente toma de conciencia de la importancia de la GEI ha hecho que los cultivares sean evaluados en múltiples ambientes y regiones para su recomendación o para su ingreso en programas de selección de materiales elite. Kang (2002) discute las implicancias de la GEI en un programa de mejoramiento genético, a saber: (a) se dificulta la identificación de materiales superiores, ya que podrá existir cambio de rankings de genotipos y (b) se incrementan los costos de evaluación, debido a que la prueba debe realizarse en varios lugares representativos de áreas de cultivo claves.

La GEI reduce la utilidad de la media genotípica a través de ambientes como indicador de la superioridad de los genotipos y disminuye la eficiencia de selección medida en términos de recursos invertidos por unidad de ganancia genética (Ceretta y Abadie, 2003). A su vez los autores (citando a Cooper y Byth, 1996) establecen que existen básicamente dos enfoques: uno *tradicional* donde la GEI es vista como fuente de error en la estimación de

medias, por lo que no intenta interpretar las causas de la GEI sino minimizar ese error mediante el muestreo ambiental tan amplio como sea posible. El otro enfoque *alternativo* trata la GEI como una fuente de información utilizándola en la selección por adaptación específica. De esta forma se hace necesario estudiar (entre otros) el grado de relación/similitud entre ambientes y caracterizarlos en términos de estrés ambiental relevante.

Por lo tanto, el estudio de la GEI involucra muchos aspectos de la ciencia biológica, fundamentalmente sobre el crecimiento y desarrollo vegetal. Una extensa proporción de la investigación en ciencia agronómica concierne al estudio de la GEI (Xu, 2010). Sin embargo se está lejos de entender los factores que influencian la adaptación, aun considerando solo los principales cultivos agrícolas.

1.2 MÉTODOS DE ANÁLISIS DE LA GEI

En su revisión sobre los métodos estadísticos para el análisis de la interacción genotipo-ambiente, Freeman (1973) relata los primeros trabajos sobre el tema, que preceden incluso al análisis de varianza. Cita el trabajo de Fisher y Mackenzie (1923) como primera referencia, postulando que el estudio de la GEI podría modelarse con un operador multiplicativo. En la misma revisión se citan trabajos que fundamentan la importancia de considerar términos de interacción en el modelo de análisis de genotipos y ambientes. Particularmente, es citado el trabajo de Morley Jones y Mather (1958) que establecen como la GEI influencia varianzas y covarianzas utilizadas para cuantificar la variación en ciertos modelos genéticos. Cuando existe GEI entonces “las medidas del efecto genético se aplican solo al rango de ambientes estudiados”.

El principal objetivo de un programa de mejora genética es la de evaluar la aptitud y conveniencia de determinados genotipos con propósitos agrícolas, a través de un rango de condicionantes agro-ecológicas (Xu, 2010). Con ese motivo se conducen ensayos multi-ambientes (del inglés, *multi-environment trial* o MET) donde estos genotipos se evalúan buscando representar un ambiente tipo u objetivo (“*target environment”*) donde seleccionar genotipos de amplia adaptación o aquellos de adaptación específica.

Según Crossa (1990) los ensayos MET tienen tres objetivos principales: (a) estimar y predecir de forma precisa los rendimientos de los individuos evaluados para formar el próximo ciclo de selección, basados en información experimental limitada, (b) determinar patrones de respuesta de genotipos o tratamientos agrícolas a través de diferentes ambientes y (c) proveer de guía fehaciente en la selección de genotipos o tratamientos agrícolas para plantar en años futuros o nuevas localidades. Evaluar un genotipo o tratamiento agronómico sin incluir su interacción con el ambiente es un análisis incompleto y por ende limita la precisión de las estimaciones de rendimiento. Por lo tanto, una parte significativa de los recursos de un programa de mejoramiento de cultivos y agronómico se dedican a determinar ésta interacción a través de ensayos repetidos en varios ambientes.

La consecuencia fundamental para el fitomejorador es que cuanto más se manifieste el componente GEI por sobre el valor genotípico, menor heredabilidad para el carácter se obtendrá en el proceso de selección y por ende, mayor es la dificultad de su mejora. El abordaje del genetista entonces, es muy diferente al del agrónomo o fitomejorador: éstos desean minimizar el efecto de la GEI en los ensayos de campo, mientras que los primeros el de entender las causas de la interacción en términos de parámetros genéticos (Freeman, 1973).

El análisis estadístico entonces debe proveer estimación de parámetros que nos indiquen cuán bien se comportan en promedio ciertos genotipos a través de un rango de ambientes, y cuán bien se comportan en condiciones específicas (Xu, 2010). Por lo tanto, en ensayos MET, la GEI se considera ausente si todos los genotipos se comportan de manera similar a través de todos los ambientes, es decir, la variación total es explicada mayormente por efectos principales de ambiente y genotipo.

1.2.1 El modelo básico

Desde un punto de vista estadístico, considerando como ejemplo el rendimiento (fenotipo observado) de una serie de genotipos evaluados en diferentes ambientes (cada uno con *r* repeticiones), puede ser representado de la siguiente manera (Bernardo, 2010)

$$y_{ijr} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ijr} \qquad [1]$$

donde y_{ijr} es la respuesta empírica medida para el *i*-ésimo genotipo (*i* = 1, 2, …, *I*) en el *j*-ésimo ambiente (*j* = 1, 2, ..., *J*) con *r* repeticiones en cada *I* x *J* combinación, μ es la media general, g_i es el efecto genotípico aditivo del *i*-ésimo genotipo; e_j es el efecto ambiental aditivo del *j*-ésimo ambiente; $(ge)_{ij}$ es la respuesta particular del *i*-ésimo genotipo en el *j*-ésimo ambiente; ε_{ijr} simboliza el error de observación o residual experimental asociado al *i*-ésimo genotipo del *j*-ésimo ambiente y *r*-ésima repetición.

El modelo en [1] es llamado modelo lineal básico, de efectos fijos a dos vías y que puede resumirse en una tabla de doble-entrada, con genotipos y ambientes en sus filas / columnas. Cada celda de esa tabla entonces se corresponde con el valor de Y_{ij} ó respuesta empírica media del fenotipo.

En ese contexto, el efecto de la GEI es entonces estimado como

$$(ge)_{ij} = Y_{ij} - g_i - e_j - \mu$$

De ésta forma, el componente no aditivo de la fórmula en [1] implica que el valor esperado de Y_{ij} depende no solo de los valores de g_i y e_j separadamente sino también la particular combinación de ellos.

Su limitación principal es que la varianza del error a través de ambientes se supone homogénea para probar las diferencias genotípicas. A su vez, el análisis de varianza combinado de múltiples ambientes no explora posibles estructuras subyacentes del componente de GEI, no determina patrones de respuesta de genotipos y ambientes. Este modelo es apropiado sólo para el análisis de medias en situaciones de ensayos balanceados (replicados equitativamente) y homocedasticidad de los errores (Crossa, 1990).

Como antecedente del problema, Finlay y Wilkinson (1963) presentaron formalmente un modelo de regresión para modelar el término de interacción considerando las ideas originales Fisher y Mackenzie (1923) y Yates y Cochran (1938).

Básicamente la idea consiste en un modelo de regresión del desempeño de los genotipos sobre los rendimientos medios de ambientes, expresados de la forma:

$$(\mathrm{ge})_{ij} = b_i e_j + d_{ij}$$

Así la interacción es descompuesta en un componente debido a la regresión lineal b_i, del *i*-ésimo genotipo en el promedio ó índice ambiental y una desviación d_{ij}.

En la práctica se ha reportado que este modelo capta una pequeña parte de la GEI. En Crossa (1990) se citan referencias donde se concluye que el método de regresión lineal es de poco valor en situaciones comunes de cultivo. Además se ha establecido que gran parte de la variabilidad de la GEI es debida a la heterogeneidad de varianzas.

1.2.2 El modelo lineal-bilineal

El modelo lineal-bilineal representa una versión multivariada del modelo de regresión antes citado (Kang *et al*., 2004), donde la respuesta del i-ésimo genotipo en el j-ésimo ambiente se modela: por una parte a través de efectos aditivos principales (g + e, componentes lineales), y por uno o más términos multiplicativos que explican el término de interacción (componente bilineal). Propuesto inicialmente en Gollob (1968), el modelo llamado *FANOVA* por su autor se basa en la técnica de análisis de componentes principales (PCA) y su propósito es expresar la variabilidad en un número reducido de nuevas coordenadas, esperando que describan dimensiones valederas de los datos originales. El modelo linear-bilineal puede expresar de forma general como:

$$\mathrm{Y}_{ij} = \mu + \mathrm{g}_i + \mathrm{e}_j + \left(\sum_{k=1}^{K} \lambda_k u_{ik}\, v_{jk}\right) + \delta_{ij} \qquad [2]$$

donde el término multiplicativo de la GEI se representa como la suma de *K* factores bilineales constituidos por el producto de un coeficiente λ_k , y los componentes (o scores) u_{ik} y v_{jk} de genotipo y ambiente, respectivamente; siendo el término δ_{ij} la parte de la interacción no captada por los *K* factores del modelo. Una forma conveniente de obtener los valores λ_k , u_{ik} y v_{jk} es a través de la descomposición del valor singular de la tabla de efectos GE. De ésta

manera, siendo H la matriz de efectos GE, a través de éste método obtenemos $H = U\Lambda V'$ donde $\Lambda = diag\{\lambda_k\}$ es la matriz de valores propios, $U = \{u_{ik}\}$ y $V = \{v_{jk}\}$ las matrices de los vectores singulares izquierdo y derecho de H (Meyer, 2009).

El modelo FANOVA es renombrado por Zobel *et al*. (1988) como AMMI (del inglés: *Additive Main Effects and Multiplicative Interaction*), posteriormente popularizado para el análisis de ensayos MET de manera rutinaria. Tanto en el citado trabajo como en subsiguientes (Gauch y Zobel 1996, Crossa 1990; Gauch 1988) se comparó el modelo AMMI con otros modelos multiplicativos y se resaltaron los beneficios de su utilización.

El modelo AMMI en [2] junto con otros modelos lineales-bilineales se describen en van Eeuwijk (1995) y Cornelius *et al*. (1996) como una metodología general de análisis MET, que pueden ser utilizados en el estudio de patrones de respuesta de genotipos y ambientes, con la ventaja de poder visualizar esos patrones gráficamente a través de un *biplot* (Gabriel, 1971).

En Crossa y Cornelius (2002) exponen los principales modelos lineales-bilineales derivados del AMMI, dedicando especial atención el modelo SREG (*Sites Regression Model*). Teniendo en cuenta que el modelo AMMI es incapaz de distinguir entre la interacción con cambio o sin cambio de ranking de genotipos, los autores sugieren que SREG es un mejor modelo para identificar grupos de ambientes donde la GEI con cambio de ranking de genotipos es despreciable. Es el modelo recomendado para mejoradores ya que los términos multiplicativos del modelo contienen el efecto principal del genotipo junto con el de interacción, de la siguiente manera:

$$Y_{ij} = \mu_j + (\sum_{k=1}^{K} \lambda_k u_{ik}\, v_{jk}) + \delta_{ij} \qquad [3]$$

donde $\mu_j = \mu + e_j$ es la media ambiental.

Más recientemente, varias revisiones sobre el tema han marcado ventajas y desventajas de éstos modelos de efectos fijos (Crossa 2012, Yang *et al*. 2009, Gauch *et al*. 2008, Yan *et al*. 2007, Yan y Tinker 2005).

1.2.3 El modelo lineal mixto

Los modelos de análisis anteriormente señalados asumen que tanto genotipos como ambientes son factores fijos. Sin embargo, en varias situaciones es más razonable considerar los efectos de los factores del modelo como una muestra aleatoria de una población. Bajo el enfoque de modelos mixtos, algunos efectos se asumen provenientes de una distribución de efectos aleatorios. Esto implica que existe (al menos conceptualmente) una más amplia población de efectos genéticos y que el muestreo nos brinda valores que son realizaciones de esa población, valores que pueden ser predichos utilizando los mejores predictores lineales insesgados (BLUP, *Best Linear Unbiased Predictor*) propuestos en Henderson (1975).

Según Xu (2010) todos los términos en [1] son usualmente aleatorios excepto la media general. Para tener una interpretación genética de la variación observada podemos considerar la variación fenotípica como la combinación de: una "señal genética" al componente [g_i + (ge) $_{ij}$] y un "contexto ambiental" al componente e_j , sumado al "ruido ambiental" del componente ε_{ij} del modelo.

Estos componentes pueden ser tomados como fijos dependiendo del método de muestreo y el propósito del análisis. Así, el ambiente puede referirse a sitios ó localidades que se consideran fijos cuando no son escogidos aleatoriamente de todos los posibles sitios en un área particular. Sin embargo, si el ambiente refiere a años entonces podemos tomarlo como efecto aleatorio, lo mismo si tratáramos con una combinación de año por localidad.

El modelo general lineal mixto es en su forma matricial:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Zu} + \varepsilon \qquad [4]$$

donde $\mathrm{y}^{(n \times 1)}$ es el vector de observaciones, $\mathrm{X}^{(n \times p)}$ y $\mathrm{Z}^{(n \times q)}$ son matrices conocidas de diseño del modelo, $\beta^{(p \times 1)}$ el vector de parámetros fijos del modelo y $\mathrm{u}^{(q \times 1)}$ y $\varepsilon^{(n \times 1)}$ son los vectores de efectos aleatorios, con distribución conjunta gaussiana, E(u) y E(ε) usualmente se asumen igual a cero y varianza

$$\mathrm{V}\begin{pmatrix}\mathrm{u}\\ \varepsilon\end{pmatrix} = \begin{pmatrix}\mathrm{G} & 0\\ 0 & \mathrm{R}\end{pmatrix}$$

Los supuestos en relación a la matriz G y R definen el modelo mixto particular de análisis. La distribución de los datos se asume normal, con $E(y)= X\beta$ y $V(y)= ZGZ' + R$ (McLean *et al*. 1991, Harville 1977, Henderson 1975).

En la medida que las matrices G y R son (o se asumen) conocidas, las estimaciones y predicciones de los efectos fijos y aleatorios pueden obtenerse simultáneamente a través del método de mínimos cuadrados generalizados mediante ecuaciones lineales mixtas (Henderson, 1975), obteniéndose los mejores estimadores lineales insesgados (BLUE, *Best Linear Unbiased Estimator*) para el vector β y los BLUPs para el vector u.

Si las matrices G y R no son conocidas, los componentes de varianza del modelo pueden estimarse a través de otras metodologías, dentro de las cuales el procedimiento de máxima verosimilitud restringida REML presenta ventajas apreciables (Corbeil y Searle 1976, Patterson y Thompson 1971)

Piepho (1997) propone ajustar un modelo mixto que contenga términos multiplicativos, que no necesariamente se asuma homogeneidad de varianzas sino modelizar la estructura de covarianza. Bajo el enfoque de modelos mixtos, los efectos aleatorios multiplicativos de GEI pueden incluir correlaciones entre las interacciones por lo que presentan una estrecha relación con estructuras de varianzas-covarianzas de factores analíticos propuesta en el trabajo de Jennrich y Schluchter (1986).

El mismo autor (Piepho, 1998) presenta un modelo multiplicativo de factores analíticos, con efectos aleatorios de genotipo y genotipo-ambiente que es conceptual y funcionalmente más apropiado al modelo fijo tipo AMMI.

En ese mismo contexto, Smith *et al*. (2001) presentan modelos multiplicativos como una clase más general capaz de proveer un enfoque más realista al análisis de datos multiambiente. Se presenta un modelo de análisis mixto que considera heterogeneidad de varianzas para la GEI, correlaciones entre las GEI e incluye además estructuras apropiadas de covarianza para ensayos individuales (correcciones espaciales para cada ambiente).

Crossa *et al*. (2006) afirma que la principal característica de la metodología de modelos mixtos en comparación a la de modelos fijos es que permite modelar estructuras de varianza-covarianza heterogéneas y correlacionadas, y muestra como utilizar diferentes modelos mixtos (para los efectos principales de genotipos y genotipo-ambiente) utilizando información de parentesco entre genotipos, mejorando la predicción de los valores genéticos de datos multiambiente.

1.3 ESTRUCTURAS DE VARIANZA

Existen diferentes estructuras que modelan la matriz de correlaciones y/o de varianzas-covarianzas y pueden ser utilizadas según la situación particular de interés a modelar. En el caso de la metodología de modelos mixtos, son diversas las situaciones en las que permite modelar estructuras de varianza.

1.3.1 Definición de estructuras de varianza

Existe una amplia gama de estructuras de varianza, tanto para modelos de correlación como de varianzas-covarianzas (ya que ambos modelos están relacionados entre sí como se observa en las fórmulas siguientes), y cada una de ellas será de utilidad dependiendo de la situación que se pretenda modelizar.

Un *modelo de correlación* es descrito de la forma

$$\mathrm{C} = \{\mathrm{c_{ij}}\} : \begin{cases} \mathrm{c_{ii}} = 1 \ \forall \mathrm{i} \\ \mathrm{c_{ij}} = \mathrm{c_{ji}} \quad |\mathrm{c_{ij}}| < 1, i \neq j \end{cases}$$

En el caso de un *modelo de varianzas de tipo homogéneas*, su representación es la siguiente

$$\mathrm{V} = \{\mathrm{v_{ij}}\} : \begin{cases} \mathrm{v_{ii}} = \sigma^2 \ \forall \mathrm{i} \\ \mathrm{v_{ij}} = \mathrm{v_{ji}} \quad \mathrm{i} \neq \mathrm{j} \end{cases} \rightarrow \mathrm{V} = \sigma^2 \mathrm{C}$$

Si el *modelo de varianza es de tipo heterogéneo*, entonces

$$V = \{v_{ij}\} : \begin{cases} v_{ii} = \sigma_i^2 & i = 1, \dots, n \\ v_{ij} = v_{ji} & i \neq j \end{cases} \rightarrow \begin{matrix} V = DCD \\ D^{n \times n} = diag(\sigma_i) \end{matrix}$$

A modo de síntesis, se presentan esquemáticamente algunas estructuras de uso frecuente en el análisis de datos MET, a saber: *simple*, de *componentes de varianza*, *simetría compuesta (homo/heterogénea)*, *autoregresivo (homo/heterogénea)*, *no estructurada* y de *factores analíticos*.

De todas las posibilidades, la más restrictiva es la estructura simple que asume que la varianza es común y constante, con covarianzas iguales a cero. En el contexto del análisis de ensayos de genotipos en múltiples sitios, las varianzas dentro de sitios se asumen iguales y las covarianzas iguales a cero, bajo el supuesto que los sitios son independientes. Esto es restrictivo en la medida que por ejemplo, un sitio en un año particular probablemente muestre mayor variación genotípica que otros sitios, y el desempeño en distintos sitios puede presentar diferentes grados de correlación para un año ó de un año a otro (Crossa *et al.*, 2004).

Otra estructura es la de componentes de varianza que permite diferentes varianzas para *k* efectos aunque asumiendo covarianzas iguales a cero:

$$\sigma_{ij} = {\sigma_k}^2 \, 1(i = j)$$

$$\begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & & \sigma_k^2 \end{bmatrix}$$

siendo $1(i = j)$ una función indicadora que asigna el valor 1 si i = j y cero lo contrario.

Si el modelo asume una varianza y covarianza constante entonces el modelo es llamado de simetría compuesta que se esquematiza seguidamente:

$$\sigma_{ij} = \sigma_1 \ + \sigma^2 \, 1(i = j)$$

$$\begin{bmatrix} \sigma_1 + \sigma^2 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 + \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 + \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 + \sigma^2 \end{bmatrix}$$

donde el número de parámetros a estimar son 2: una varianza y covarianza constante. Esta estructura se corresponde también con la llamada de correlación uniforme, ya que se descompone en un parámetro de varianza y un parámetro de correlación únicos a estimar.

En el caso de varianzas heterogéneas, es decir diferentes varianzas en los elementos de la diagonal (σ_i^2) y correlación constante, la estructura será de simetría compuesta heterogénea.

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$$

El modelo autoregresivo de primer grado posee una varianza común, constante pero la covarianza declina exponencialmente, lo que puede aplicarse a situaciones donde mediciones cercanas en el tiempo o el espacio están correlacionadas.

$$\sigma_{ij} = \sigma^2 \rho^{|i-j|}$$

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

En el caso de varianzas heterogéneas, entonces la estructura se denomina autoregresivo heterogéneo:

$$\sigma_{ij} = \sigma_i \sigma_j \rho^{|i-j|}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$$

En el otro extremo, tenemos la estructura más liberal, llamada no estructurada, que permite a cada término de la matriz ser diferente. Deben por lo tanto, estimarse un número de parámetros igual a *t*(*t*+1)/2 siendo *t* el orden de la matriz puesto que $\sigma_{ij} = \sigma_{ji}$. En el caso de un modelo mixto, ésta estructura de varianza-covarianza para la matriz G es demasiado general para revelar patrones, por lo que un modelo más parsimonioso es deseable (Crossa *et al.*, 2004).

$$\sigma_{ij} = \sigma_{ij}$$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

1.3.2 Estructura de factores analíticos

La estructura de *factores analíticos* brinda una solución intermedia entre estructuras restrictivas y no restrictivas. La misma modela la covarianza en términos de factores latentes (no observados) lo que puede ser útil para modelar la matriz G del modelo mixto:

$$\sigma_{ij} = \sum_{k=1}^{\min(i,j,q)} \lambda_{ik}\,\lambda_{jk} + {\sigma_i}^2 1(i=j)$$

$$\begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1q} \\ \lambda_{21} & \cdots & \lambda_{2q} \\ \vdots & \ddots & \vdots \\ \lambda_{t1} & \cdots & \lambda_{tq} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{t1} \\ \vdots & & \ddots & \vdots \\ \lambda_{1q} & \lambda_{2q} & \cdots & \lambda_{tq} \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & \cdots & & 0 \\ & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sigma_t^2 \end{bmatrix}$$

que especifica un modelo de factores analíticos con *q* factores (Jennrich y Schluchter 1986) de forma $\Lambda\Lambda' + \Psi$, donde $\Lambda^{(t\ x\ q)}$ es una matriz de coeficientes (loadings) y $\Psi^{(t\ x\ t)}$ una matriz diagonal de varianzas específicas. El número de parámetros va a depender entonces

del número *q* de factores elegido, tal que para el modelo general FA(*q*) el número es igual a [*q*/2](2*t* -*q* + 1) + *t*, siendo *t* la dimensión de la matriz de covarianza.

A partir del trabajo de Piepho (1998) se ha despertado el interés por la aplicación del modelo FA en el contexto de modelos mixtos, particularmente para la modelación de la varianza genotípica y la GEI.

En Smith *et al*. (2001) se asume que la estructura de varianza para el término aleatorio del modelo tiene forma separable de modo que

$$\mathrm{var}(\mathrm{u_g}) = \mathrm{G_e} \otimes \mathrm{G_v}$$

donde $\mathrm{G_e}$ y $\mathrm{G_v}$ son matrices simétricas de *p* y *m* dimensiones correspondientes al componente de ambiente y variedad, respectivamente. Por simplicidad se asume que $\mathrm{G_v} = \mathrm{I_m}$ y la matriz $\mathrm{G_e}$ es la llamada matriz de varianza genética. A su vez, la misma puede estar dada por:

$$\mathrm{G_e} = \sigma_{\mathrm{v}}^2 \, \mathrm{J_p} + \sigma_{\mathrm{ve}}^2 \, \mathrm{I_p}$$

siendo la matriz de varianza explicada por un componente del efecto de la variedad más la interacción variedad-ambiente, y $\mathrm{J_p}$ una matriz de unos para *p* ambientes, generando así la estructura de simetría compuesta que como ya fue mencionado, raramente provee un adecuado ajuste a los datos. En el citado trabajo, Smith *et al*. consideran la estructura de factores analíticos al efecto de variedad para cada ambiente.

En Burgueño *et al*. (2007) se utilizó el modelo de factores analíticos siguiendo el modelo en [3], donde el efecto aleatorio del *i*-ésimo genotipo en el *j*-ésimo ambiente se describe como una función linear (sin intercepto) de variables latentes x_{ik} con coeficientes λ_{jk} para k=1,2,…,K más un residual δ_{ij}:

$$g_{ij} = \sum_{k=1}^{K} \lambda_{jk} \, x_{ik} + \delta_{ij}$$

donde λ_{jk} es el *loading* del j-ésimo ambiente (expresa la potencialidad ambiental) en el k-ésimo factor latente, x_{ik} es el score del i-ésimo genotipo (expresa sensibilidad genotípica) en el k-ésimo factor latente, y δ_{ij} es el término del error no explicado del modelo.

En su forma matricial, g_{ij} puede ser expresado como:

$$\mathrm{g} = (\Lambda \otimes \mathrm{I_g})\mathrm{x} + \delta$$

Asumiendo que los genotipos no están relacionados, la matriz de varianzas-covarianzas de g es

$$\mathrm{G} = (\Lambda\Lambda^{'} + \Psi) \otimes \mathrm{I_g}$$

donde para una estructura de factores analíticos de k factores [FA(k)] con $k \leq s$, $\Lambda^{(s \times k)}$ es una matriz de λ y $\Psi^{(s \times s)}$ una matriz diagonal de *s* elementos diferentes. El componente genético del modelo está dado por la matriz de identidad $\mathrm{I_g}$, que asume independencia entre genotipos. Como resultado del producto kronecker ($\otimes$) se obtiene una matriz diagonal en bloques donde en cada bloque se disponen las matrices de varianzas-covarianzas genotípicas. En el caso de tener disponible información de parentesco entre genotipos, ésta puede ser incluida en el modelo (ver como ejemplo Beeck *et al.* 2010, Kelly 2009, Meyer 2009, Crossa *et al*. 2006).

De esta forma, puede interpretarse la estructura FA como una regresión lineal del genotipo y genotipo-ambiente sobre la covariable ambiental (loadings ambientales), donde cada genotipo posee una pendiente particular que mide la sensibilidad de los genotipos al factor latente ambiental representado por el loading de cada ambiente.

1.4 ANÁLISIS CON DATOS DESBALANCEADOS

En ensayos MET, las pruebas de cultivares generalmente presentan situaciones de desbalance, no necesariamente debido a la pérdida no planificada de parcelas o falta de semilla. De los diversos motivos, es común que algunos genotipos sean probados sólo en

algunos sitios o años y que algunos otros ingresen (nuevos materiales a evaluar) o sean descartados del sistema de evaluación. Como consecuencia del proceso selectivo de cultivares se obtiene un conjunto de datos altamente desbalanceado, con un importante número de celdas vacías para los datos de año por lugar por cultivar (Crossa, 1990).

En los programas de mejoramiento genético la metodología de análisis de varianza (ANOVA) ha sido históricamente la utilizada para la estimación de componentes de varianza, relacionado a diversas fuentes de variación, incluida la GEI. El método de ANOVA brinda siempre estimaciones insesgadas, aunque permite más de una definición de suma de cuadrados (secuencial vs parcial) por la que se debe optar. También pueden obtenerse estimaciones mediante métodos basados en la verosimilitud, como el método de máxima verosimilitud (ML) y máxima verosimilitud restringida (REML). Estos métodos son preferidos porque permiten el tratamiento de datos desbalanceados o estructuras de datos complejas (Littell, 2002), aunque se cita en la literatura que las estimaciones por ML no son insesgadas. Con el método REML se supera ésta desventaja y produce idénticos estimadores que ANOVA, siempre y cuando los datos sean balanceados y las estimaciones no negativas (Yang 2007, Crossa 1990). Una revisión completa sobre el tema fue propuesta inicialmente por Hartley y Hocking (1971) y Rubin (1976).

En Zobel *et al*. (1988) los autores compararon diferentes estrategias de análisis para estimar el efecto del genotipo y la interacción con el ambiente. Para ello utilizaron ensayos de rendimiento en soja consistente en más de 70 variedades pero no todas fueron testeadas en todos los años y sitios; el análisis se limitó a considerar 7 variedades en 35 ambientes (combinación año-sitio). Además, para la mayoría de los ensayos se poseían 4 repeticiones, pero de las 980 potenciales parcelas, algunos sólo contaban con dos o tres repeticiones debido a pérdidas no planificadas. Según los autores, el análisis se basó en fórmulas de cálculo del error más simples y de menor esfuerzo computacional, primando el sentido práctico ya que “la estrategia alternativa de descartar casi la mitad de las observaciones para obtener un diseño balanceado no era atractivo”.

Gauch y Zobel (1990) al considerar el modelo AMMI como un modelo de efectos fijos trataron el tema de datos faltantes, marcando la limitación práctica que ocasionan en el

análisis. Los autores implementaron el algoritmo EM para ajustar un modelo EM-AMMI, modelo que toma en cuenta datos faltantes. Este considera (citando a Searle, 1987) información directa proporcionada por las *R* repeticiones del experimento de sembrar *G* genotipos en E ambientes, y la información indirecta que aporta el resto de los *GER* datos, es decir *GER-R*. Según los autores, cada una de las observaciones del experimento tiene influencia sobre cada uno de los parámetros del modelo, por lo que no solo las repeticiones afectan las estimaciones, sino también las restantes *GER-R* repeticiones.

En su revisión, Kang (1997) trata el tema del desbalance de datos en el análisis de la GEI desde el punto de vista de la estabilidad de varianzas de Shukla (1972). Según el autor citando a Searle (1987), usualmente el efecto del ambiente es tratado como aleatorio y el efecto de cultivar como fijo (por ende la interacción entre ambos es aleatoria) la inferencia sobre los efectos aleatorios del modelo mediante mínimos cuadrados no es apropiada cuando tenemos desbalance en los datos. El hecho de que podamos tener estimaciones negativas de la varianzas es otro punto débil de este enfoque. Por lo que recomienda el uso de un modelo mixto y máxima verosimilitud en la estimación de la varianza de la GEI. Como demostración de procedimientos, el autor analiza un ensayo de rendimientos en maíz (11 híbridos en 4 localidades y 4 años) que por diversos motivos era desbalanceado, aunque no se dispone del nivel de desbalance. Mediante estimación REML se obtuvieron los BLUEs y componentes de varianza en el sentido de Shukla. Como conclusión, el cálculo de varianzas mediante REML para datos desbalanceados permite obtener mejores estimaciones de los parámetros de estabilidad, además de superar las dificultades del desbalance en los datos. Piepho y Mohring (2006) exploraron patrones de datos faltantes comunes en esquemas de evaluación de cultivares, y las implicancias sobre los efectos del cultivar y componentes de varianza, basados en la verosimilitud.

Los resultados de dicho estudio revelaron que es deseable usar la mayor cantidad de información posible en la estimación de componentes de varianza (e.g. usando varios años). El análisis basado en el modelo con efecto principal del cultivar aleatorio proporciona estimaciones más precisas de la GEI que el análisis donde el efecto principal del cultivar se modela como fijo. Esto es debido, según los autores, a las propiedades asintóticas de los métodos basados en la verosimilitud.

En Roozeboom *et al*. (2008) se analizaron 21 años de ensayos de rendimiento en trigo (de un total de 102 genotipos y 17 localidades) mediante REML para la estimación de componentes de varianza, y biplots para delinear localidades en grupos conteniendo genotipos con el mismo patrón de rendimiento. Según los autores los datos eran altamente desbalanceados, fundamentalmente entre años. El número de genotipos cambia cada año ya que nuevos materiales son introducidos, mientras que aquellos obsoletos son eliminados progresivamente. En cuanto a las localidades, no fueron consistentes todos los años, debido a que algunas se perdieron por razones extremas (sequía, inundación, congelamiento) o de manejo (problema de plagas no controladas). Obligados por esta situación, el trabajo consistió en la comparación entre años de los datos balanceados a un mínimo de 9 genotipos en 10 localidades, lo que le resta relevancia al estudio a nivel del mejoramiento genético.

En su trabajo sobre imputación de datos faltantes en ensayos de interacción genotipo-ambiente, Arciniegas-Alarcón *et al*. (2010) simularon varias tasas de pérdida de datos aleatoria (10, 20 y 40%) sobre un conjunto de datos real. El objetivo del trabajo fue el de proponer un algoritmo determinístico de imputación de datos, mediante el método de validación cruzada. Luego, estos métodos de imputación fueron probados sobre otro conjunto de datos diferente, al que se eliminó aleatoriamente un 30% de sus datos. Mediante ANOVA, los efectos principales del modelo fueron estimados luego de la imputación, y se compararon las diferencias con respecto a la matriz completa original.

Burgueño *et al*. (2011) comparan la habilidad predictiva de los modelos lineales mixtos cuando la GEI se modela a través de factores analíticos, mediante un esquema de validación cruzada que elimina aleatoriamente algunos genotipos de algunos sitios. Destacan las ventajas del modelo lineal mixto en lo que respecta a la facilidad para manejar datos incompletos. A pesar de que sus objetivos no se relacionan directamente con el presente estudio, se desprende que su metodología fue apta para evaluar y predecir el comportamiento de genotipos no observados en sitios y/o años, que es en definitiva el principal objetivo del mejorador. Seis METs reales fueron evaluados: 1 en papa, 3 en maíz y 2 en trigo que incluyen sitios y genotipos para el primer y segundo caso, mientras que para el último incluyó sitios, genotipos y años. Para éstos datos fueron eliminados de forma aleatoria

de 1 a 3 genotipos para cada ambiente, estos genotipos faltantes fueron luego predichos por el modelo con datos completos.

1.5 COMPARACIÓN ENTRE ESTRUCTURAS DE VARIANZA

De la extensa bibliografía existente sobre el análisis de la GEI predomina aquella cuyos objetivos son resumir la variabilidad y encontrar patrones que permiten una suerte de agrupamiento entre sitios y/o genotipos, reduciendo la GEI (Bernardo 2010, Kang 2002, Crossa 1990). En general estas metodologías son clasificadas como multivariadas, tal es el caso de PCA, *Cluster Analysis*, *Factor Analysis*, *CoA*, etc.

En Flores *et al*. (1998) se compararon 22 metodologías univariadas (paramétricas y no paramétricas) y multivariadas, mostrando las ventajas y desventajas de su uso desde la perspectiva del agrónomo y mejorador genético. A su vez, en diversos trabajos se ejemplifican diversos enfoques integradores (Cullis *et al*. 2010, Burgueño *et al*., 2008, Yan y Tinker, 2005) cuyo objetivo es modelar asociaciones entre sitios, agrupamiento de genotipos y formación de megambientes.

Con el propósito de comparar las diferentes estructuras de varianza, en este estudio se proponen dos herramientas estadísticas muy diferentes entre sí, en la metodología que utilizan para medir la similaridad entre matrices o estructuras de covarianza: (a) *test de Mantel* y (b) análisis de *Procrustes*. Ambas pueden ser utilizadas para la comparación y evaluación de la semejanza entre dos o más matrices basadas en los mismos objetos.

Para el tratamiento teórico y aplicaciones de estas metodologías se recomienda Schneider y Borlund (2007), Jackson (1995) y Rohlf y Slice (1990).

1.5.1 Test de Mantel

Debido a su sencillez y flexibilidad, el enfoque más común para medir la similaridad (congruencia) entre dos conjuntos de datos multivariados es el test de correlación entre matrices desarrollado por Mantel. Es útil para la comparación entre dos matrices de similaridad (proximidad), y de esa forma poder afirmar si existe de forma significativa una correlación entre matrices. La forma original del estadístico de Mantel se define como la suma de los productos cruzados de los valores no estandarizados entre dos matrices de similaridad, cuando estas se disponen en dos vectores de igual orden, excluyendo los valores de la diagonal (Mantel, 1967 citado por Schneider y Borlund, 2007). Para mejorar su interpretación como medida descriptiva, se sugiere la estandarización de los valores, obteniendo de esa forma un estadístico análogo al coeficiente r de Pearson. Como tal, expresa la relación lineal que existe entre matrices de proximidad lo que asume una distribución normal. Alternativamente, puede utilizarse el estadístico de correlación de Spearman como medida no paramétrica, computándose la correlación de rangos entre los valores de proximidad entre matrices (Dietz, 1983).

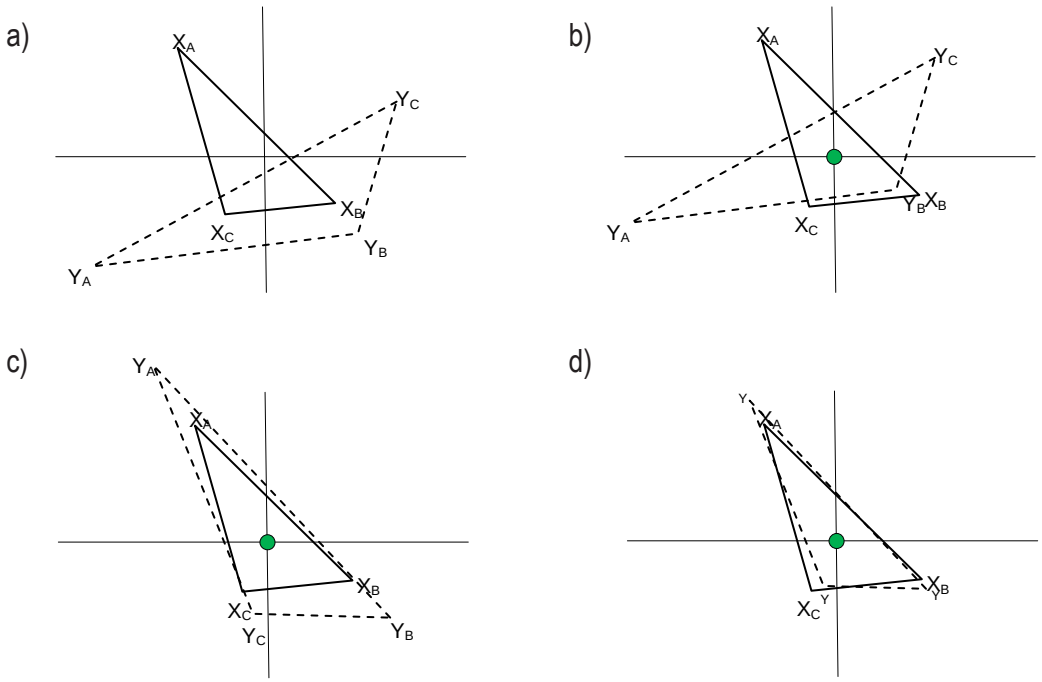
A la hora de probar la significancia del coeficiente de correlación entre matrices de proximidad se ha ideado el test de aleatorización o permutación (Manly, 2005). Es decir que, para estimar el p-valor de un estadístico de Mantel observado entre dos matrices, se somete a una de las matrices a un cambio o permutación aleatoria de sus elementos, siendo el estadístico recalculado. Repitiendo el procedimiento utilizando diferentes ordenamientos se obtienen valores aleatorios que conforman entonces la distribución de referencia del estadístico de Mantel bajo la hipótesis nula. Si la hipótesis nula de no correlación es verdadera entre dos matrices, entonces la permutación de filas y columnas producirá mayores o menores valores del estadístico con la misma probabilidad. La significancia se estima como la proporción de permutaciones que produce un coeficiente de correlación igual o mayor a la correlación original observada.

1.5.2 Análisis de Procrustes

En el análisis de Procrustes, en contraste con la técnica de Mantel, busca el parecido entre dos o más configuraciones para el mismo conjunto de objetos, que surge generalmente de (a) técnicas de proyección de ejes principales (*eigenanalysis*), caso típico del análisis de componentes principales, o (b) técnicas de optimización de las distancias entre objetos, buscando minimizar una función objetivo determinada, caso típico de la técnica de escalamiento multidimensional.

El procedimiento de superponer las coordenadas de un objeto (matriz Y) para el mejor ajuste a otra configuración de puntos de referencia (matriz X), manteniendo rígidas las posiciones relativas entre puntos de la configuración a ajustar, se conoce como análisis de Procrustes (PA). El criterio de ajuste simple es el de mínimos cuadrados, llamado análisis de Procrustes ortogonal (Gower, 1975), donde se busca minimizar la suma de cuadrados de las diferencias entre dichas configuraciones en un espacio euclidiano p-dimensional.

El procedimiento esquematizado en el siguiente diagrama, consiste en (1) *centrar* las coordenadas del primer objeto en el segundo: es decir el traslado de todos los puntos por una medida constaste en una dirección común, cambiando entonces el origen a un centroide común, (2) *escalar* ambos objetos, esto es dividir los elementos de ambas matrices por una cantidad constante de forma tal que la suma de las distancias al cuadrado de los puntos sea igual a la unidad, y por último (3) *rotar* las coordenadas de la primer configuración (Y) hacia la referencia (X) mediante el movimiento rígido de todos sus puntos en un ángulo constante (sin variación en la distancia de los puntos al centroide).



Ejemplo del Análisis de Procrustes (adaptado de Schneider y Borlund, 2007). En a) se muestra dos configuraciones simples de puntos (triangulo X e Y) de diferente tamaño, donde X es la forma objetivo ó referencia y Y la forma a ajustar. En b) se presenta el efecto de la configuración luego del traslado a un centro común (efecto traslación), c) luego de ser rotado (efecto rotación) y d) reducido por un factor constante (efecto escala) de forma que el criterio de Procrustes sea optimizado.

Otra forma de presentar el método de Procrustes es

$$\text{matriz } \mathbf{X} \text{ referencia} = \text{transformación}(\text{ matriz } \mathbf{Y}) + \text{residual}$$

donde la transformación incluye los siguientes efectos

$$\mathbf{X} = \left\{ \begin{matrix} efecto \\ translación \end{matrix} \right\} + \left\{ \begin{matrix} efecto \\ escala \end{matrix} \right\} \times \mathbf{Y} \times \left\{ \begin{matrix} efecto \\ rotación \end{matrix} \right\} + \text{residual}$$

Matricialmente, puede ser reescrito

$$\mathbf{X} = 1\, c^{'} + \rho\, \mathbf{Y}\, \mathrm{A} + e$$

donde las matrices $\mathbf{X}$ e $\mathbf{Y}$ de tamaño $p\ x\ k$ son matriz de referencia y matriz a ajustar, respectivamente, c es un vector fila (de translación) de k elementos, ρ es un escalar (factor de escala) y A es la matriz de rotación (ortogonal) de tamaño $k\ x\ k$ tal que $\mathrm{A'A} = \mathrm{I}$; y el vector e de k elementos se corresponde con los residuales del modelo.

Por lo que el análisis de Procrustes consiste en obtener estimaciones de los parámetros mencionados resolviendo un problema de minimización matricial

$$\mathbf{min} \left\| \mathbf{X} - \left(1\ c^{'} + \rho\ \mathbf{Y}\ \mathrm{A} \right) \right\|$$

Por lo tanto, luego del ajuste de Procrustes se obtiene una matriz transformada que minimiza la suma de cuadrados de las distancias entre los puntos de ambas configuraciones, donde el estadístico de Procrustes m^2 mide la bondad de ajuste resultante de una configuración de referencia con otra de interés, proveyendo una comparación válida entre ellas.

Krzanowski (1987) utilizó el PA para medir la discrepancia entre una configuración (matriz completa) de referencia comparado con configuraciones de la misma matriz modificada. En su trabajo sobre selección de variables que preserven una estructura definida mediante la técnica de componentes principales, obtuvo un indicador de cual era la pérdida de información en la estructura de los datos, a medida que se eliminan variables de la matriz completa.

En Jackson (1995) se utilizó el PA para buscar la concordancia ecológica de abundancia de comunidades de invertebrados en ciertos lagos, su morfología y química así como su localización geográfica. A través de un análisis de correspondencia y de componentes principales, el estudio busca la comparación de diversos indicadores ecológicos y su correspondencia con la localización geográfica de los lagos. Luego, a través de un test de permutaciones se evaluó la probabilidad de que una permutación aleatoria de esos puntos obtenga un score de similaridad mayor que la observada en los datos. El test fue llamado PROTEST y se encuentra actualmente implementado en el paquete estadístico de R.

Peres-Neto y Jackson (2001) ilustran la aplicación de Procrustes (y concretamente del gráfico superimpuesto) para examinar visualmente la concordancia de las observaciones para cada dimensión separadamente, lo que nos provee una mejor interpretación de la estructura en los datos. Dicho trabajo contrasta la efectividad en términos de potencia y de tasa de error tipo I del test de Mantel y el análisis de Procrustes, señalando las múltiples ventajas de este último.

Más cercano en el tiempo Wang *et al*. (2010) compararon espacialmente mapas de población-genéticos humanos mediante el PA. En dicho trabajo se compararon mapas genéticos de poblaciones en diferentes países de Europa contra coordenadas geográficas de una muestra de localidades. Como resultado de dicho análisis, y confirmado por un test de permutaciones, demuestran que existe un alto nivel de similaridad entre el biplot PCA de la información genética y el mapa geográfico, concluyendo la muy alta concordancia entre la genética y la localización geográfica.

Surge de la información consultada que ambas técnicas, el test de Mantel y el análisis Procrustes están muy relacionados, aunque para el primero el grado de parecido entre dos matrices está sujeto a la elección de una medida de distancia de las muchas existentes. Alternativamente, en el caso de Procrustes se comparan los datos originales (o los que surgen como coordenadas o componentes principales). A su vez, mientras que con el test de Mantel obtenemos un estadístico que comprende las diferencias de las observaciones globalmente, con Procrustes se puede evaluar cada variable o dimensión individualmente y directamente de forma gráfica. Ambos se valen de un test de aleatorización o permutaciones para evaluar si el valor de correlación obtenido es diferente al obtenido por azar.

En síntesis, la información proveniente de ensayos multiambiente generalmente es i) altamente desbalanceada (un cultivar particular es observado en un sub-conjunto de todos los ambientes) por selección o por falta circunstancial del dato, ii) es frecuente la presencia de GEI, especialmente cuando ésta es del tipo compleja (con cambio en el ranking de cultivares) y iii) el supuesto de homogeneidad de varianza de los efectos aleatorios no se sustenta.

El enfoque de modelos mixtos para el análisis estadístico de los datos con las características antes mencionadas provee un marco de trabajo más flexible ya que no depende, hasta un cierto límite, de información completa, y es posible modelar cuales efectos son fijos y cuales aleatorios permitiendo a éstos últimos diferentes situaciones de homogeneidad/heterogeneidad a través de una estructura de varianzas-covarianzas más conveniente. De las diferentes estructuras de varianza, la estructura de factores analíticos permite modelizar de forma parsimoniosa la variación GEI.

En la siguiente sección se presenta el trabajo experimental en formato de artículo científico, y que tiene como objetivo cuantificar el efecto en la estructura de varianza de la GEI cuando ésta es modelada bajo un enfoque de modelo mixto y estructura de factores analíticos, a partir de la simulación de diferentes grados y tipos de pérdida aleatoria de datos, bajo la hipótesis de que la presencia de datos faltantes tiene un efecto importante y cuantificable en el análisis de la GEI. El efecto del desbalance dependería del grado de pérdida (% de datos) en que se ven afectados nuestros datos, así como el tipo de información que se pierde (pérdida en parcelas, genotipos y/o sitios).

2. EFECTO DE DATOS FALTANTES EN LA INTERPRETACIÓN DE LA INTERACCIÓN GENOTIPO-AMBIENTE: UN ENFOQUE DE MODELO MIXTO CON FACTORES ANALÍTICOS §.

Víctor Prieto[1], Juan Burgueño[2]

1: Depto. de Biometría, Estadística y Computación, Facultad de Agronomía. Avda. E. Garzón 780. Cód. Postal 12900 Montevideo, Uruguay. Correo electrónico: vprieto@fagro.edu.uy

2: Biometrics and Statistics Unit, Crop Informatics Lab. (CRIL), International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico, D.F., Mexico

2.1 RESUMEN

Como consecuencia del proceso selectivo de cultivares se obtiene un conjunto de datos altamente desbalanceado, con un importante número de celdas faltantes para los datos de año x localidad x cultivar. En los ensayos de cultivares muchos materiales son descartados para años subsiguientes mientras que otros nuevos se incorporan. Estos datos pueden ser analizados bajo un enfoque de modelos lineales mixtos posibilitando así el tratamiento de observaciones faltantes de forma directa y modelizando una estructura de covarianza de factores analíticos (FA) para la interacción genotipo por ambiente (GEI). El presente trabajo parte del objetivo de analizar la estructura de covarianza e interpretar la GEI a través de la simulación de dos esquemas de pérdida aleatoria: de parcelas y de genotipos. Fueron utilizados datos de rendimiento en grano de 19 ambientes y 49 genotipos provenientes de ensayos de trigo semi-árido de CIMMYT. La comparación entre estructuras de covarianza se basó en la correlación de Mantel y ajuste de Procrustes. Surgen como resultados una consistente pérdida de similaridad entre la matriz completa y la que presenta desbalance, se pierde el patrón de respuesta de los materiales y se manifiestan efectos de escala en la variabilidad residual del ajuste de Procrustes. A su vez, estos cambios se diferencian en magnitud dependiendo de la estructura de varianza: correlación, varianzas-covarianzas o de

§ El formato de las citas y referencias bibliográficas del presente capítulo se adecua a lo requerido en la revista Agrociencia Uruguay.

la propia estructura FA. Los resultados del presente trabajo permitirían a los mejoradores tomar decisiones respecto a la confiabilidad que puedan presentar resultados de sus análisis dependiendo de la pérdida de información que tengan.

Palabras clave: datos faltantes**,** modelo mixto de factores analíticos, interacción genotipo-ambiente, Procrustes.

Effect of missing data in the interpretation of genotype-environment interaction: a factor analytic mixed model approach.

2.2 SUMMARY

As a result of the selection process of cultivars within a plant breeding program, a highly unbalanced datasets are obtained with a large number of missing cells for year x location x cultivar. In cultivar trials many of the materials are discarded for subsequent years while new ones are incorporated. These data can be analyzed under the mixed linear model approach enabling the treatment of missing observations directly and also modeling a factor analytic (FA) covariance structure of genotype by environment interaction (GEI). The aim of the present study is to analyze the covariance structure and interpretation of GEI through simulation of two random different data loss schemes: on plots and on genotypes. A semi-arid CIMMYT wheat yield trial was used consisting of 49 genotypes and 19 international environments. The comparison between the variance structures was based on Mantel test correlation and Procrustes analysis. Arise as a result of this work a consistent loss of similarity between the complete matrix and with one presenting missing data, differences in the response pattern of the materials, and manifesting scale effects on the residual variability of Procrustes fitted model. In addition, these evidenced changes differ in magnitude depending on the structure of variation analyzed: correlation or variance-covariance or the FA structure. The results of this study could allow researchers to make decisions about the reliability of their results that may arise depending on the information they have lost.

Keywords: unbalanced data, factor analytic mixed model, genotype by environment interaction, Procrustes.

2.3 INTRODUCCIÓN

En el mejoramiento de cultivares un número extenso de genotipos son normalmente evaluados sobre un amplio rango de ambientes, que incluye diferentes sitios, años, épocas de siembra y prácticas de cultivo, entre otros. Esto se debe a la diferente expresión de los genotipos según el ambiente, lo que se conoce como interacción genotipo-ambiente (GEI, *Genotype by Environment Interaction*). Esta respuesta diferencial medida en valores observables (fenotipo) como componentes del rendimiento, altura de planta, etc. puede ocasionar incluso el cambio de posición relativa o ranking de cultivares para diferentes ambientes (Annichiarichio 2002, Kang y Gauch, 1996). La evaluación de un genotipo o de cualquier tratamiento agronómico sin incluir su interacción con el ambiente limita la precisión en la estimación de su rendimiento, por consiguiente se destina una significativa proporción de los recursos en los programas de mejora para determinar esa interacción, a través de ensayos con múltiples ambientes (Crossa, 1990).

La implicancia fundamental para el fitomejorador es que cuanto más se manifieste el componente GEI por sobre el valor genotípico, menor heredabilidad para el carácter se obtendrá en el proceso de selección y por ende, mayor es la dificultad en su mejora. Otras implicancias determinantes para un programa de mejoramiento genético son que (a) se dificulta la identificación de materiales superiores, ya que podría existir cambio de rankings de genotipos, lo que se conoce como COI (*crossover interaction*) y (b) se incrementan los costos de evaluación, debido a que la prueba debe realizarse en varios lugares representativos de áreas de cultivo claves (Kang, 2002).

En ensayos con múltiples ambientes (MET, *multi-environment trials*), las pruebas de cultivares pueden presentar diferentes situaciones de desbalance, incluso dictados por el

diseño experimental elegido (diseños incompletos). Pero las causas por las cuales el desbalance no planificado puede ser importante no necesariamente se debe a la pérdida imprevista de parcelas que pueden surgir en el transcurso de un experimento. En etapas iniciales del mejoramiento por ejemplo, es la disponibilidad de semilla la que limita la capacidad de prueba de todos los materiales. En cambio, en etapas posteriores es común que algunos genotipos sean descartados por tener malos rendimientos, o que se mantengan otros por tener estabilidad a través de los años, y que ingresen nuevos materiales al sistema de evaluación. Como consecuencia de este proceso selectivo de cultivares se obtiene un conjunto de datos altamente desbalanceado, con un importante número de celdas vacías para los datos de año por sitio por genotipo (Crossa, 1990). En este sentido, toma importancia la utilización de metodologías que maximicen la información que se posee, que permita obtener estimaciones precisas dentro del contexto mencionado, y así utilizar mejor los recursos del programa de prueba.

En los programas de mejoramiento genético la metodología de análisis de varianza (ANOVA) ha sido intensamente utilizada para la estimación de componentes de varianza, relacionado a diversas fuentes de variación, incluida la GEI. El método de ANOVA brinda siempre estimaciones insesgadas, aunque permite más de una definición de suma de cuadrados por la que se debe optar. También pueden obtenerse estimaciones mediante métodos basados en la verosimilitud, como el método de máxima verosimilitud (ML) y máxima verosimilitud restringida (REML). Estos métodos son los preferidos porque permiten el tratamiento de datos desbalanceados o estructuras complejas de datos, aunque se cita en la literatura que las estimaciones por ML no son insesgadas. Con el método REML se supera ésta desventaja y produce idénticos estimadores que ANOVA, siempre y cuando los datos sean balanceados y

las estimaciones no negativas (Crossa 1990, Freeman, 1973). Para el caso donde exista desbalance en los datos, es posible el análisis bajo el enfoque de modelos lineales mixtos y estimación REML, que posibilita el tratamiento de observaciones faltantes de forma directa (Kelly *et al*, 2007, Balzarini 2002). En su trabajo de simulación con datos altamente desbalanceados, Piepho y Mohring (2006) concluye que la estimación REML es preferible a ML debido al menor sesgo y error. Recomienda usar toda la información posible a través de estudios de más largo plazo y además, indica que el análisis donde el efecto del genotipo es tratado como aleatorio brinda mejores estimaciones de componentes de varianza de la GEI que tratarlos como efectos fijos.

Es en ese contexto de análisis donde existen diversas maneras de modelizar la matriz de varianza genética, que van desde el supuesto de que todos los ambientes tienen la misma varianza genética (y misma covarianza para cada par de ambientes distintos) hasta la estructura más general donde todos los elementos de la matriz se permiten diferentes (no estructurada). En el análisis de datos de mejoramiento genético, estructuras más parsimoniosas son más eficientes ya que permiten modelar correlaciones genéticas y entre observaciones con un número menor de parámetros a estimar (Balzarini 2002, Smith *et al.* 2001, Piepho 1997). Dentro de estas alternativas ha tomado interés el modelo de factores analíticos (*FA, factor analytic*) ya que no solo permite una solución más parsimoniosa para modelar la estructura de covarianza de la GEI, sino que además, al estar asociado a un modelo mixto que supone efectos aleatorios permite manejar datos desbalanceados de forma directa (Burgueño *et al.* 2007, Crossa *et al.* 2006). En su trabajo sobre imputación de datos faltantes en ensayos de interacción genotipo-ambiente, Arciniegas-Alarcón *et al*. (2010) simularon varias tasas de pérdida de datos aleatoria (10, 20 y 40%) sobre un conjunto de

datos real. El objetivo del trabajo fue el de proponer un algoritmo determinístico de imputación de datos, mediante el método de validación cruzada. Luego, éstos métodos de imputación fueron probados sobre otro conjunto de datos diferente, al que se eliminó aleatoriamente un 30% de sus datos. Mediante ANOVA, los efectos principales del modelo fueron estimados luego de la imputación, y se compararon las diferencias con respecto a la matriz completa original. No surge del trabajo el análisis sobre el efecto multiplicativo de interacción de los datos faltantes. Más recientemente, Yan (2013) reportó un procedimiento para estimar datos faltantes a través del método de descomposición de valor singular (SVD) y análisis de gráficos tipo biplot. Mediante simulación, la estimación fue exitosa cuando el porcentaje de datos faltantes no superó el 40% y el conjunto de datos es pequeño (18 cultivares y 9 ambientes). Cuando el conjunto de datos es mayor (15 cultivares y 40 ambientes) la estimación resulta exitosa aún cuando el 60% de los datos fue tratada como faltante.

El presente trabajo tiene como objetivo el análisis del rendimiento de ensayos multiambiente provenientes de un programa de evaluación de cultivares de trigo, simulando diferentes grados de pérdida de datos aleatoria. Este análisis, bajo un enfoque de modelos mixtos y estimación REML, modelando el componente GEI a través de una estructura de factores analíticos permitió comparar el efecto de diferentes niveles de pérdida aleatoria de datos, a través del análisis de Procrustes. A su vez, mejora la interpretación de la GEI mostrando en cuales sitios o genotipos es mayor/menor el efecto de la pérdida de datos.

2.4 MATERIALES Y MÉTODOS

Datos experimentales.

Para llevar a cabo el trabajo se utilizaron datos de rendimiento en grano (Mg ha^{-1}) de 49 genotipos de trigo [*Triticum aestivum L.*] provenientes de ensayos SAWYT (*Semi-Arid Wheat Yield Trial*) del Centro Internacional para el Mejoramiento de Maíz y Trigo (CIMMYT) para 19 localidades internacionales de siembra, analizados bajo un diseño experimental de bloques completos al azar con 2 replicaciones por sitio.

Generación de datos faltantes.

Esta matriz completa fue sometida a diferentes niveles de pérdida de datos simuladas por un proceso de muestreo al azar simple sin remplazo. Se definieron dos tipos de pérdida aleatoria (ver Cuadro 1): (a) pérdida de parcelas, que va desde un 5 a 50% de datos eliminados

Cuadro 1. Tipos de pérdida aleatoria simulada: (a) porcentaje de pérdida de parcelas y (b) número de genotipos perdidos por sitio.

Tipo de pérdida		
	(a) % de parcelas	(b) Nº de genotipos
1)	5	3
2)	10	5
3)	15	7
4)	20	10
5)	25	12
6)	30	15
7)	35	17
8)	40	20
9)	45	22
10)	50	25
Total de celdas = 19 sitios x 49 genotipos x 2 rep = 1862		

y (b) pérdida de genotipos dentro de sitios, que va desde 3 a 25 genotipos eliminados para cada uno de los 19 sitios. Para cada nivel de pérdida se obtuvieron 50 muestras aleatorias que hacen un total de 1000 muestras.

Modelo de análisis

El análisis de los datos sigue el modelo SREG de efectos mixtos

$$Y_{ijk} = \mu + s_j + r_{k(j)} + \sum_{k=1}^{q}\left(\lambda_{jk}x_{ik} + \delta_{ij}\right) + e_{ijk} \quad [1]$$

donde Y_{ijk} es la variable de respuesta, μ la media general, s_j el efecto fijo del j-ésimo sitio (*j* = 1,...,s), $r_{k(j)}$ el efecto de la replicación dentro de sitios y e_{ijk} el error experimental. El efecto principal de genotipo y la interacción genotipo por ambiente se modela conjuntamente como una función lineal de x_{ik} de variables latentes y coeficientes λ_{jk} más un residual δ_{ij} , siguiendo a Burgueño *et al.* (2007). El componente λ_{jk} es el loading del j-ésimo sitio sobre el k-ésimo factor latente, el x_{ik} el score del i-ésimo genotipo sobre el k-ésimo factor latente y δ_{ij} el error no explicado por ésta estructura de factores analíticos.

Para mayor simplicidad, el modelo expresado de forma matricial es el siguiente:

$$y = X\beta + Zg + \varepsilon \quad [2]$$

siendo y el vector de observaciones, X y Z las matrices de incidencia para los efectos fijos y aleatorios respectivamente, β el vector de efectos fijos, g y ε los vectores de efectos aleatorios y del error respectivamente, que se asumen normal e independientemente distribuidos, con E(g) y E(ε) igual a cero y varianzas-covarianzas

$$\mathrm{V}\begin{pmatrix} g \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \mathrm{G} & 0 \\ 0 & \mathrm{R} \end{pmatrix}$$

siendo la matriz de varianzas-covarianzas de *y* igual a V(y)= $Z\mathrm{G}Z'+\mathrm{R}$ (Harville, 1977). Los supuestos en relación a la matriz G y a la matriz R definen el modelo mixto particular de análisis. Para el caso de los errores, la matriz R tiene la forma simple $\sigma_e^2 \otimes I_n$ dado por el

producto kronecker de la varianza residual y la matriz identidad de orden *n* = (*s* × *g* × *r)*, asumiendo que no existe correlación entre parcelas de un mismo ambiente.

Para la matriz G, que consta del componente genotipo + genotipo × sitio, adquiere la forma separable

$$\mathrm{G} = \boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_g = \quad [3]$$

$$\begin{bmatrix} \sigma^2_{g_1} & \rho_{12}\sigma_{g_1}\sigma_{g_2} & \rho_{1s}\sigma_{g_1}\sigma_{g_s} \\ \rho_{21}\sigma_{g_2}\sigma_{g_1} & \sigma^2_{g_2} & \\ & & \\ \rho_{s1}\sigma_{g_s}\sigma_{g_1} & & \sigma^2_{g_s} \end{bmatrix} \otimes \Sigma_g$$

ó producto kronecker entre las varianzas-covarianzas Σ_s y Σ_g, siendo éstas los componentes ambiental y genética de la matriz G. Con respecto al primero, comprende una matriz de dimensión s x s conteniendo las varianzas genéticas dentro de sitios en su diagonal, y el resto de los elementos la covarianza genética $\rho_{ij}\sigma_{g_i}\sigma_{g_j}$ entre los sitios i y j, siendo ρ_{ij} la correlación de los efectos genéticos entre cada par de sitios (Crossa *et al.*, 2004). El componente genético de G, Σ_g, es una matriz de identidad de dimensión *g*, lo que presupone independencia entre los genotipos.

En este trabajo el componente ambiental de G, Σ_s, es modelado mediante una estructura de factores analíticos, por lo tanto

$$\mathrm{G} = (\Lambda\Lambda' + \Psi) \otimes I_g \quad [4]$$

$$\left(\begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{s1} & \lambda_{s2} & \cdots & \lambda_{sq} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{sq} \\ \lambda_{12} & \lambda_{22} & \ldots & \lambda_{s2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1q} & \lambda_{2q} & \cdots & \lambda_{sq} \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \psi_s \end{bmatrix} \otimes I_g \right)$$

donde Λ es una matriz de orden $s \times q$ ($q \leq s$) cuyas columnas contienen los loadings ambientales del k-ésimo factor o covariable latente (k = 1,…, q) y Ψ la matriz de varianzas específicas para cada sitio, indicando falta de ajuste al modelo factorial. Cuando un solo factor es considerado, entonces k=1 y el modelo se denota FA(1), si son dos el modelo será FA(2), y así sucesivamente. Existen varios estudios que indican que el modelo FA(2) es el más indicado en el sentido de que el agregado de más factores no redunda en una mayor precisión de predicción del modelo (Burgueño *et al*. 2011).

En el presente trabajo los loadings de ambientes y scores de genotipos obtenidos de cada análisis fueron rotados siguiendo el método de componentes principales, es decir, el primer factor cuenta con máxima varianza, el segundo le sigue en la variabilidad que capta y ortogonal al primero, y así sucesivamente (Smith *et al.* 2005).

Comparación de resultados.

Para la descripción e interpretación de los modelos se valió de gráficos de tipo *biplot*, que surgen del análisis del modelo FA(2). De dicho análisis se obtuvieron loadings ambientales que captan la potencialidad ambiental, lo que permite observar patrones de rendimiento similares entre los diferentes materiales evaluados. La comparación entre las matrices de varianzas-covarianzas (de tamaño s x s) de los datos completos contra aquellas con los diferentes niveles de datos faltantes se basó en el estadístico de correlación no paramétrica ρ (test de Mantel) y el análisis de Procrustes (PA). En el primer caso el objetivo es evaluar si existe una correlación significativa entre dos matrices de similaridad (Manly, 2005). En el segundo caso, la comparación es a través del estadístico m^2 de Procrustes siendo un indicador válido de la concordancia/similaridad entre matrices (Peres-Neto y Jackson 2001, Jackson D. A. 1995). Para ambos análisis fueron utilizadas diferentes librerías implementadas

en el paquete R (R Development Core Team, 2009) y para el ajuste de los modelos lineales mixtos fue utilizado el software ASReml (VSN International, 2010); su código puede ser solicitado a los autores.

2.5 RESULTADOS

Descripción para datos completos

En la Figura 1 se observa cómo se manifiestan los diferentes patrones de respuesta a través de un gráfico biplot, donde en un mismo par de ejes se disponen los loadings ambientales junto con los scores genéticos para el modelo de datos completo. En el gráfico es posible detectar aquellos sitios extremos que se ubican más distantes del origen como el caso de S14, S13, S19, S7 y S15 que se asocian con mayor variabilidad GE. De éstos, el sitio S14 y S13 presentan variabilidad debido a cambio de ranking (COI) a diferencia de S19 que es sin cambio de ranking (ver Crossa *et al*., 2004). Puede observarse como se agrupan los sitios S10, S16, S18 y S17 [cuadrante sup. der del origen], así como los sitios S11, S6, S8 y S5 [cuadrante sup. izq del origen], ambos grupos de rendimientos promedio y baja COI, con escasa correlación entre ellos. Aparecen los sitios S12 y S19 con una muy buena relación aunque este último de mayor rendimiento. El sitio S15 tiene correlación negativa con los sitios S14 y S7, encontrándose en cuadrantes opuestos del biplot.

En cuanto a genotipos, aunque su delineamiento en grupos no es tan claro, vemos que genotipos como el 20, 40, 43, 18, 48 [cuadrante inf. der. del origen] tienden a una interacción negativa con la mayor parte de los sitios, especialmente para los sitios S14 y S7, mientras que los genotipos 29, 45, 13 y 5 presentan una interacción positiva con el S14. Los genotipos 44 y 11 se observan con una alta relación positiva para el sitio S13, y negativa con el sitio

S15, mientras que los genotipos 31, 33 y 35 para el sitio S13 son de tipo negativa. Es esperable que genotipos con scores extremos presenten importante variabilidad debido a COI con varios de los sitios ya mencionados, tal es el caso entre los genotipos 8, 38, 44 y 11, y los sitios S13, S14 y S15. El análisis de los gráficos biplot es una herramienta visual útil para inspeccionar la variabilidad GE.

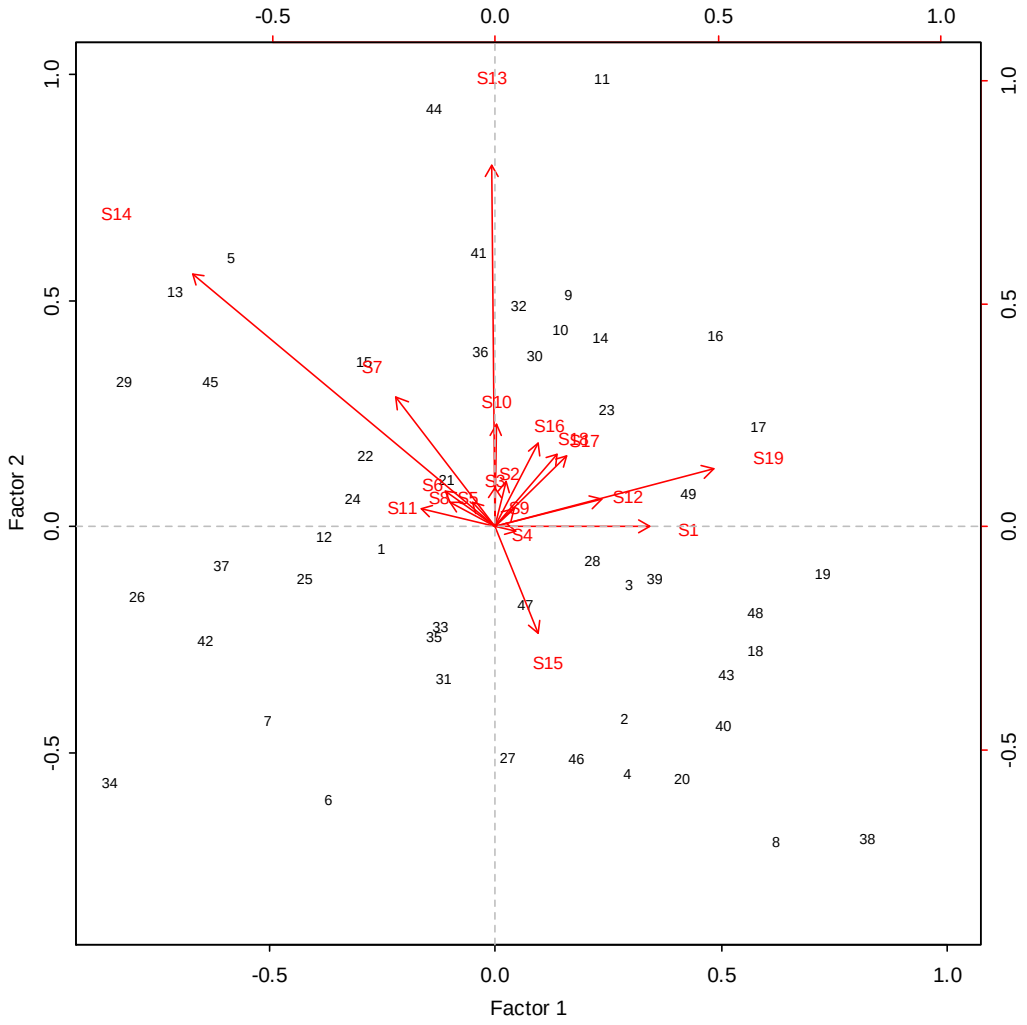


Figura 1 Biplot del modelo FA(2) con datos completos, incluyendo 19 sitios (S1-S19) y 49 genotipos (1-49).

En este trabajo se hace necesario utilizar otra técnica que nos brinde un indicador de similaridad para la comparación entre todas las muestras analizadas.

Descripción para pérdida aleatoria de parcelas

Con el fin de presentar un estadístico que resuma la similaridad entre las estructuras de varianza para los diferentes % de pérdida de datos, se presentan los resultados del test de Mantel para las matrices de correlación y de varianzas-covarianzas. De la comparación entre matrices del modelo completo y las que surgen de la simulación se observa una consistente pérdida de similaridad, expresada en el estadístico de correlación de Mantel (Cuadro 2). Se desprende del cuadro que existen diferencias según se trate de la matriz de correlación o la de varianzas-covarianzas. En ésta última, en la comparación incluye los elementos de su diagonal, mientras que en la de correlación la diagonal es trivial (todos sus elementos son 1s) y no son considerados.

Cuadro 2. Media, Desvío estándar y Coeficiente de variación para el estadístico de correlación del test de Mantel, para matrices de correlación (CORR) y varianzas-covarianzas (VCOV) de la GEI, según el porcentaje de pérdida de parcelas simulada.

	Media		*Desvío estándar*		*Coef. de variación*	
	CORR	*VCOV*	*CORR*	*VCOV*	*CORR*	*VCOV*
Tipo de pérdida aleatoria de parcelas (% de datos perdidos)						
5	0,845	0,837	0,101	0,088	0,119	0,106
10	0,806	0,792	0,104	0,080	0,129	0,101
15	0,706	0,769	0,158	0,106	0,223	0,137
20	0,666	0,712	0,155	0,089	0,232	0,124
25	0,559	0,638	0,148	0,090	0,264	0,141
30	0,537	0,628	0,121	0,086	0,225	0,137
35	0,504	0,591	0,154	0,096	0,306	0,162
40	0,491	0,595	0,177	0,123	0,361	0,206
45	0,411	0,525	0,142	0,114	0,347	0,218
50	0,326	0,444	0,159	0,123	0,488	0,277

Puede notarse que hay una tendencia lineal de pérdida en la similaridad en la medida que aumenta el % de datos faltantes, tendencia que se manifiesta en las matrices de correlación más que la de covarianza para los mismos niveles de pérdida. A medida que aumenta el % de datos faltantes, mayor es la caída en la similaridad para las matrices de correlación.

Esa tendencia sugiere que por encima del 35 % de pérdida de datos, el coeficiente promedio de r entre el modelo original y los modelos incompletos es menor o igual a 0,50, mientras en el caso de las varianzas-covarianzas eso sucede cuando se pierde el 50 % de los datos. También se manifiesta una mayor dispersión en las aleatorizaciones con el aumento en el % de datos faltantes, manifestándose en mayor medida para las matrices de correlación.

Se presenta a continuación los resultados del PA para las mismas matrices, como también para las matrices de loadings (de tamaño 19 x 2 elementos) y la matriz de scores (de tamaño 49 x 2 elementos). En el caso de las matrices de correlación y de varianzas-covarianzas, existe un aumento en el estadístico m^2 indicando una caída en la bondad de ajuste del modelo con datos perdidos con respecto al de datos completos (Cuadro 3). Se evidencia que a mayor cantidad de datos faltantes, es menor la concordancia entre matrices, y eso se traduce en una mayor falta de ajuste. Entre ambas estructuras existen marcadas diferencias en cuanto a sus valores promedios y la dispersión entre aleatorizaciones. Para las varianzas-covarianzas los valores de m^2 indican menor falta de ajuste, manteniéndose por debajo de 0,20 para todos los valores de pérdida estudiados, mientras que para las matrices de correlación, por encima de un 15% de pérdida el promedio supera el 0,26. En cuanto a la dispersión entre aleatorizaciones, si bien para las correlaciones se manifiesta en valores de desvío estándar mayores, el coeficiente de variación se muestra similar.

Cuadro 3: Media, Desvío estándar y Coeficiente de variación para el estadístico m^2 de Procrustes, para matrices de correlación (CORR) y varianzas-covarianzas (VCOV) de la GEI, según el porcentaje de pérdida de parcelas simulada.

	Media		*Desvío estándar*		*Coef. de variación*	
	CORR	*VCOV*	*CORR*	*VCOV*	*CORR*	*VCOV*
Tipo de pérdida aleatoria de parcelas (% de datos perdidos)						
5	0,130	0,028	0,098	0,018	0,755	0,636
10	0,161	0,041	0,110	0,016	0,684	0,392
15	0,266	0,051	0,164	0,021	0,616	0,413
20	0,304	0,066	0,152	0,023	0,500	0,350
25	0,399	0,084	0,154	0,023	0,386	0,267
30	0,429	0,100	0,128	0,036	0,298	0,359
35	0,449	0,111	0,149	0,039	0,331	0,350
40	0,465	0,127	0,170	0,042	0,365	0,333
45	0,542	0,150	0,140	0,041	0,259	0,272
50	0,612	0,189	0,158	0,059	0,259	0,310

Cuadro 4: Media, Desvío estándar y Coeficiente de variación para el estadístico m^2 de Procrustes, para matrices de loadings (LOAD) y scores genéticos (SCORE) de la GEI, según el porcentaje de pérdida de parcelas simulada.

	Media		*Desvío estándar*		*Coef. de variación*	
	LOAD	*SCORE*	*LOAD*	*SCORE*	*LOAD*	*SCORE*
Tipo de pérdida aleatoria de parcelas (% de datos perdidos)						
5	0,162	0,184	0,114	0,132	0,706	0,718
10	0,212	0,240	0,113	0,120	0,535	0,498
15	0,241	0,280	0,132	0,128	0,546	0,459
20	0,299	0,321	0,135	0,119	0,451	0,370
25	0,363	0,432	0,130	0,135	0,358	0,313
30	0,392	0,445	0,152	0,141	0,387	0,316
35	0,409	0,488	0,144	0,141	0,352	0,288
40	0,395	0,493	0,150	0,129	0,381	0,261
45	0,456	0,577	0,153	0,137	0,336	0,237
50	0,517	0,633	0,157	0,129	0,303	0,204

Si el PA lo hacemos sobre los loadings ambientales y scores genéticos de la estructura FA, se observa la misma tendencia en cuanto a la pérdida de concordancia en relación al modelo de datos completo (Cuadro 4). Se verifica una falta de ajuste algo mayor para las matrices de scores, aunque la variabilidad entre aleatorizaciones no se observan diferencias.

Si comparamos todas las estructuras de variación analizadas construyendo un valor índice de valor 100 para la mínima pérdida (5 % de pérdida) podemos observar como ha sido el cambio en el estadístico m^2 a medida que aumenta la pérdida de datos (Figura 2). Se observa en dicha figura un aumento en más de 6 veces en la matriz de varianzas-covarianzas cuando se pasa de 5 a 50% de pérdida. Le siguen en magnitud las matrices de correlación, aunque recién luego del 25% de pérdida, estas estructuras se diferencian entre sí.

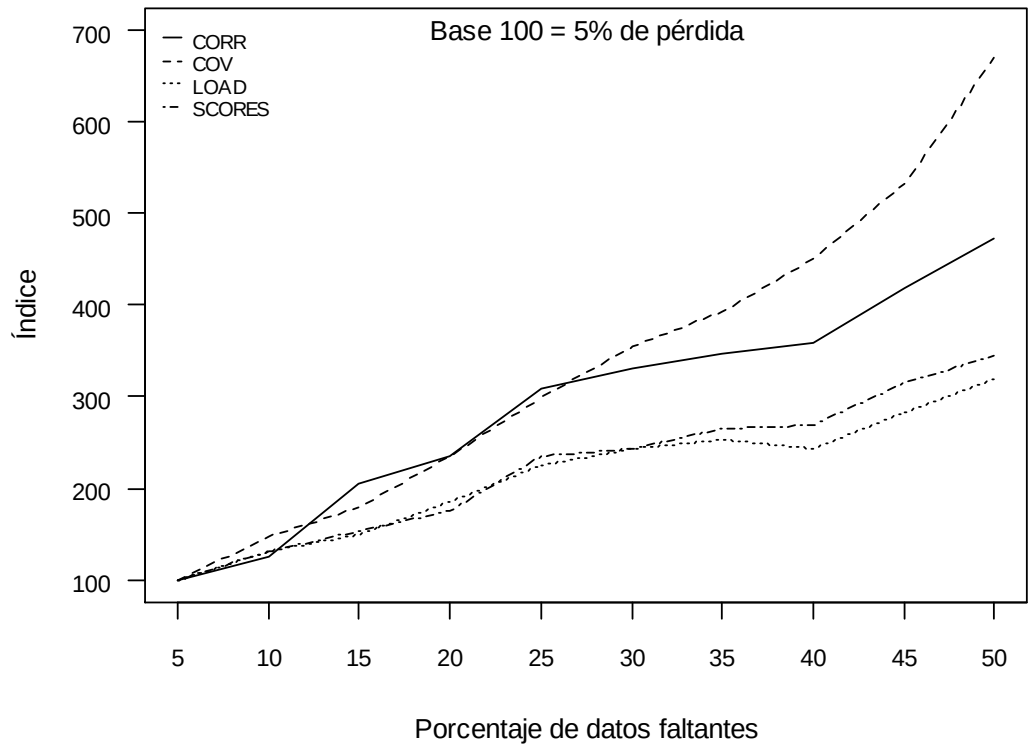


Figura 2. Índice de cambio del estadístico m^2 de Procrustes para las matrices de correlación (CORR), varianzas-covarianzas (COV), loadings ambientales (LOAD) y scores genéticos (SCORES) de la GEI, según porcentaje de datos faltantes.

Por último, las matrices de loadings y scores presentan un comportamiento similar en todo el rango de pérdida analizado, llegando a aumentar el estadístico de Procrustes más de 3 veces su valor para un 50 % de pérdida.

En cuanto al factor de escala del modelo Procrustes, puede observarse en la Figura 3 la disminución paulatina de valores y su mayor dispersión, a medida que aumenta la pérdida de datos. Los valores medios caen por debajo de 0,8 para pérdidas mayores al 25% de datos.

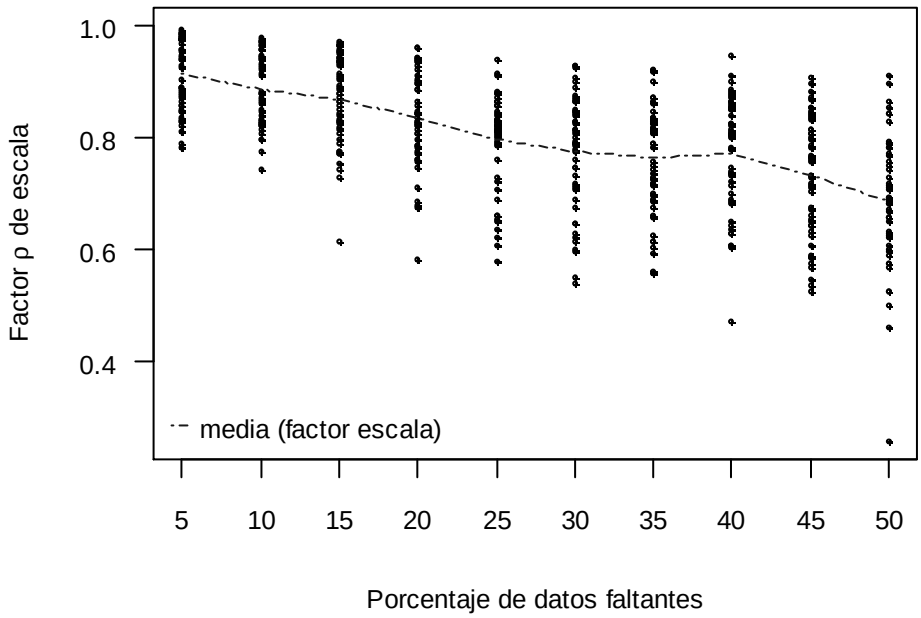


Figura 3. Factor *ρ* de escala del ajuste Procrustes para las matrices de loadings ambientales de la GEI, según porcentaje de datos faltantes.

Otra forma de analizar el ajuste del modelo del PA es mediante el estudio de sus residuales, esto es analizar los valores de *RMSE* (*Root Mean Squared Error*) que es una medida global del tamaño de los residuales. Consistentemente con lo analizado con respecto al estadístico m^2, se observa en la Figura 4 la misma tendencia en los residuales, de aumento paulatino con el aumento en el porcentaje de datos faltantes. En el gráfico se observa como la

estructura de correlación y loadings ambientales presentan mayores valores de *RMSE*; en cambio los scores genéticos y la covarianza presentan menores valores de *RMSE*, indicando entonces una menor sensibilidad al aumento en el porcentaje de datos faltantes.

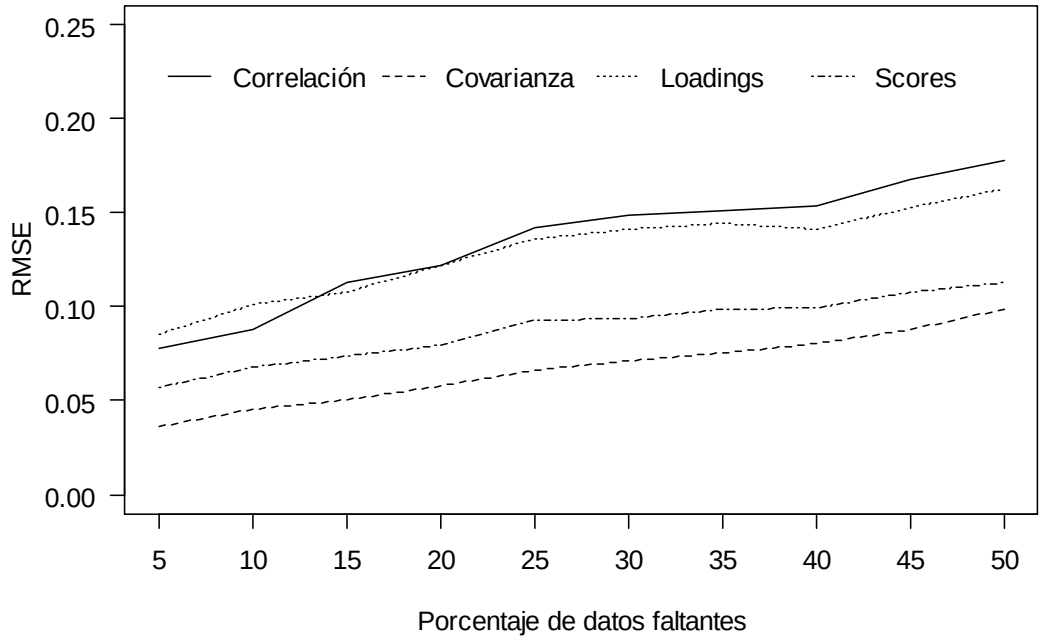


Figura 4. Valores de RMSE del ajuste de Procrustes a máxima similaridad con respecto a la matriz completa, según porcentaje de datos faltantes, para las matrices de correlación, covarianza, loadings ambientales y scores genéticos.

Si el estudio de residuales lo hacemos por sitio, cosa que es factible hacerlo en el marco del PA, puede observarse el cambio ocurrido para cada sitio en la medida en que el porcentaje de datos faltantes aumenta (Figura 5).

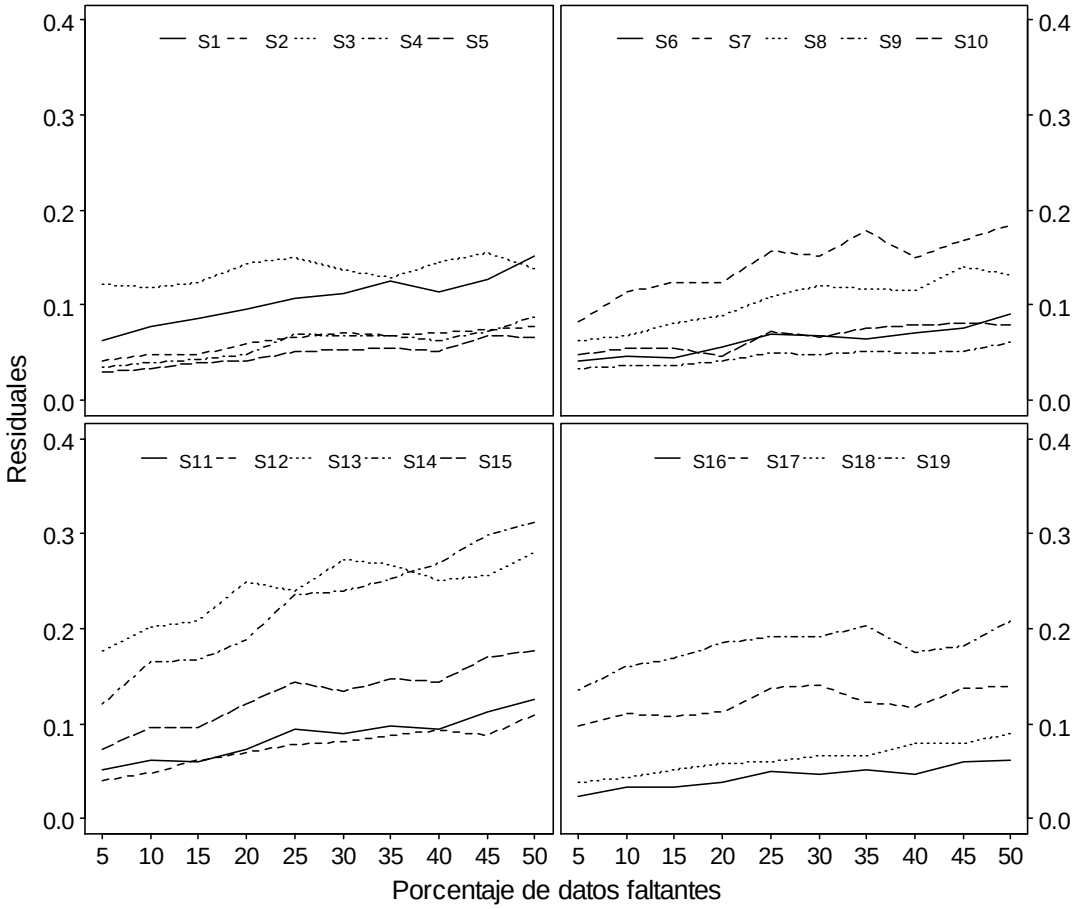


Figura 5. Valores de residuales del ajuste de Procrustes para las matrices de loadings ambiental, según porcentaje de datos faltantes, para los sitios S1 – S19.

Se observan 3 tipos de respuesta: i) el grupo de cambio *extremo*, como el caso de los sitios S13 y S14, donde para bajos porcentajes de pérdida los valores residuales superan en gran medida al resto, ii) el grupo de cambio *moderado*, que se manifiesta con aumento sistemático de los residuales, aunque se mantiene con valores por debajo de 0,20. Tal es el caso de los sitios S1, S7, S8, S11 y S15. Dentro de este grupo se agregan los sitios S3, S17 y S19 que presentan un comportamiento moderado pero errático, con altas y bajas; y por último iii) el grupo de cambio *estable* (se mantienen por debajo de 0,10), siendo poco sensibles al aumento en la pérdida de datos. Son los sitios S2, S4, S5, S6, S9, S10, S16 y S18.

Descripción para pérdida aleatoria de genotipos

En ésta sección se presentan resultados para el tipo de pérdida b), resaltando fundamentalmente aquellas diferencias con respecto al tipo de pérdida aleatoria de parcelas, anteriormente descrita. En el Cuadro 5 y 6 se describen los resultados del test de Mantel para la comparación de las matrices de correlación y varianzas-covarianzas, y las matrices de loadings y scores de la GEI. Se observa que la pérdida de genotipos provoca, al igual que en la pérdida aleatoria de parcelas, una caída importante en la similaridad entre matrices, y es mayor la caída para las matrices de correlación, lo mismo que el aumento de la dispersión en las aleatorizaciones. En el caso de loadings y scores su comportamiento es muy similar.

Se puede observar en la Figura 6 como se ve afectado el factor de escala del ajuste Procrustes. Se observa una caída pronunciada en los valores cuando se pierden más de 10 genotipos por sitio.

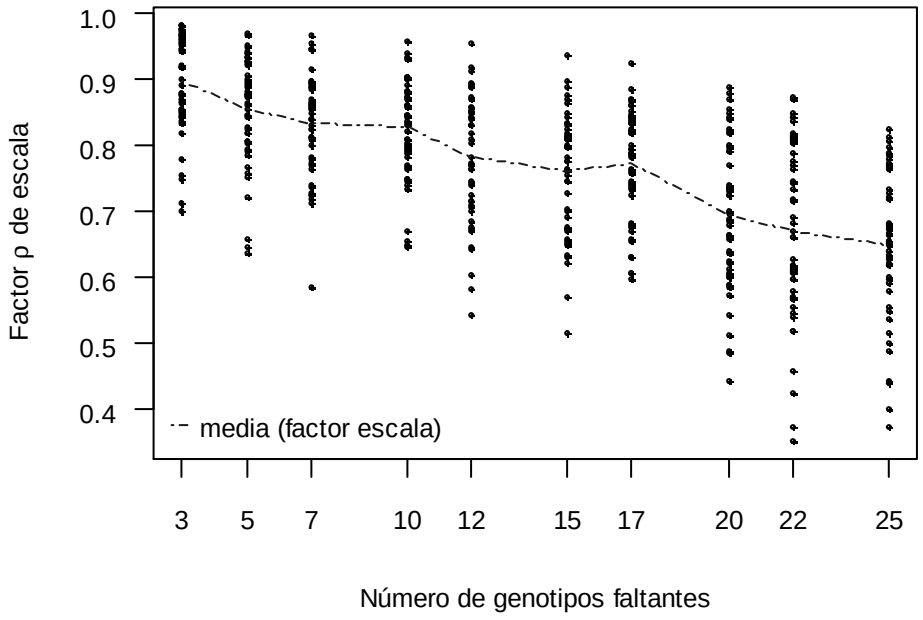


Figura 6. Factor *ρ* de escala del ajuste Procrustes para las matrices de loadings ambientales de la GEI, según número de genotipos faltantes.

Cuadro 5. Media, Desvío estándar y Coeficiente de variación para el estadístico de correlación del test de Mantel, para matrices de correlación (CORR) y varianzas-covarianzas (VCOV) de la GEI, según el número de genotipos perdidos por sitio.

	Media		*Desvío estándar*		*Coef. de variación*	
	CORR	*VCOV*	*CORR*	*VCOV*	*CORR*	*VCOV*
Tipo de pérdida aleatoria de genotipos (nº de genotipos perdidos)						
3	0,812	0,819	0,108	0,084	0,133	0,102
5	0,720	0,754	0,127	0,086	0,176	0,114
7	0,659	0,706	0,141	0,095	0,215	0,135
10	0,592	0,660	0,157	0,102	0,265	0,155
12	0,536	0,622	0,151	0,097	0,282	0,156
15	0,465	0,560	0,188	0,107	0,403	0,190
17	0,462	0,565	0,156	0,109	0,338	0,193
20	0,362	0,462	0,192	0,122	0,530	0,264
22	0,366	0,483	0,170	0,117	0,464	0,243
25	0,270	0,398	0,155	0,108	0,574	0,272

Cuadro 6: Media, Desvío estándar y Coeficiente de variación para el estadístico m^2 de Procrustes, para matrices de loadings (LOAD) y scores genéticos (SCORE) de la GEI, según el número de genotipos perdidos por sitio.

	Media		*Desvío estándar*		*Coef. de variación*	
	LOAD	*SCORE*	*LOAD*	*SCORE*	*LOAD*	*SCORE*
Tipo de pérdida aleatoria de genotipos (nº de genotipos perdidos)						
3	0,191	0,214	0,128	0,124	0,667	0,578
5	0,261	0,282	0,132	0,132	0,506	0,468
7	0,297	0,359	0,123	0,124	0,413	0,345
10	0,311	0,398	0,123	0,121	0,394	0,305
12	0,377	0,455	0,144	0,127	0,383	0,280
15	0,414	0,515	0,134	0,125	0,324	0,243
17	0,397	0,526	0,128	0,120	0,324	0,228
20	0,504	0,634	0,155	0,125	0,306	0,197
22	0,532	0,635	0,174	0,107	0,327	0,168
25	0,566	0,720	0,139	0,107	0,246	0,149

En las figura 7 se presenta como ha sido el cambio de las estructuras de variación analizadas. Se puede observar como la pérdida de genotipos afecta en mayor medida a las matrices de covarianza aumentando al doble en el estadístico de Procrustes ya con 7 genotipos faltantes, llegando a superar en 8 veces cuando se eliminan 25 de los 49 genotipos. Sin embargo, el resto de las estructuras parece tener un incremento más mesurado en el valor del estadístico al aumento del número de genotipos faltantes.

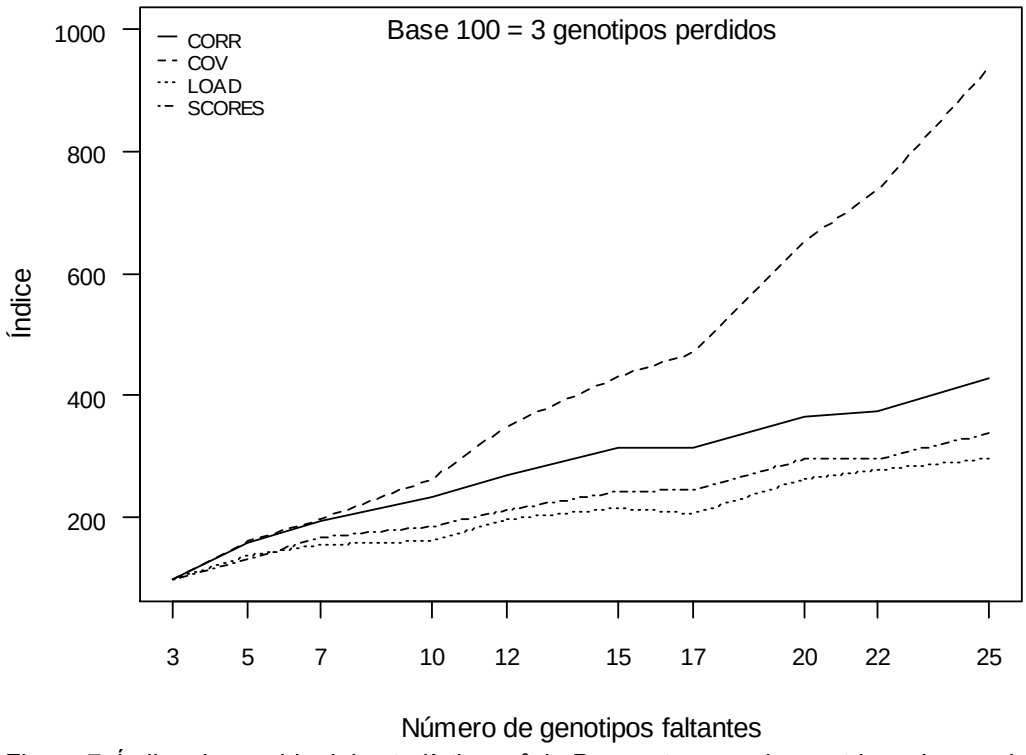


Figura 7. Índice de cambio del estadístico m^2 de Procrustes para las matrices de correlación (CORR), varianzas-covarianzas (COV), loadings ambientales (LOAD) y scores genéticos (SCORES) de la GEI, según número de genotipos faltantes.

En el caso de los residuales globales, dado por el valor de RMSE, las matrices de correlación y de loadings tienen una tendencia muy similar aún cuando las primeras mantienen valores por encima del resto (Figura 8). En el caso de las matrices de varianzas-covarianzas, presenta siempre una tendencia a su aumento con la pérdida de más genotipos por sitio.

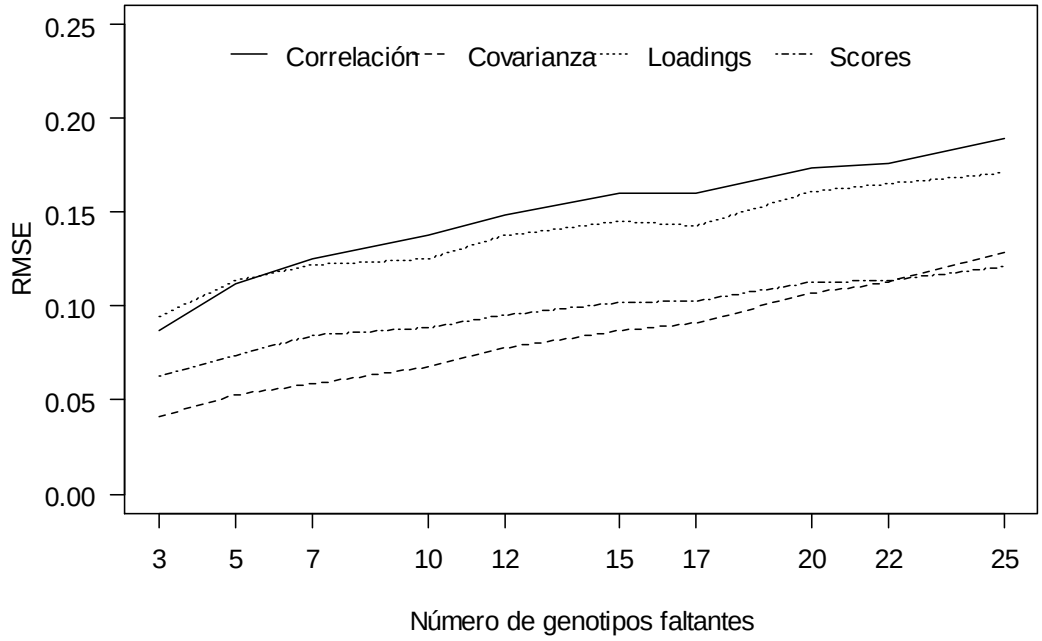


Figura 8. Valores de RMSE del ajuste de Procrustes a máxima similaridad con respecto a la matriz completa, según número de genotipos faltantes, para las matrices de correlación, covarianza, loadings ambientales y scores genéticos.

Por último, se presentan los valores de los residuales del ajuste de Procrustes (Figura 9), donde puede observarse cuáles son los sitios de mayor aporte al incremento en el estadístico m^2. Si a éstos resultados lo comparamos con la figura 4 (correspondiente a la pérdida aleatoria de parcelas) se observa que, con pocas diferencias en sus valores, muestran el mismo patrón de variación.

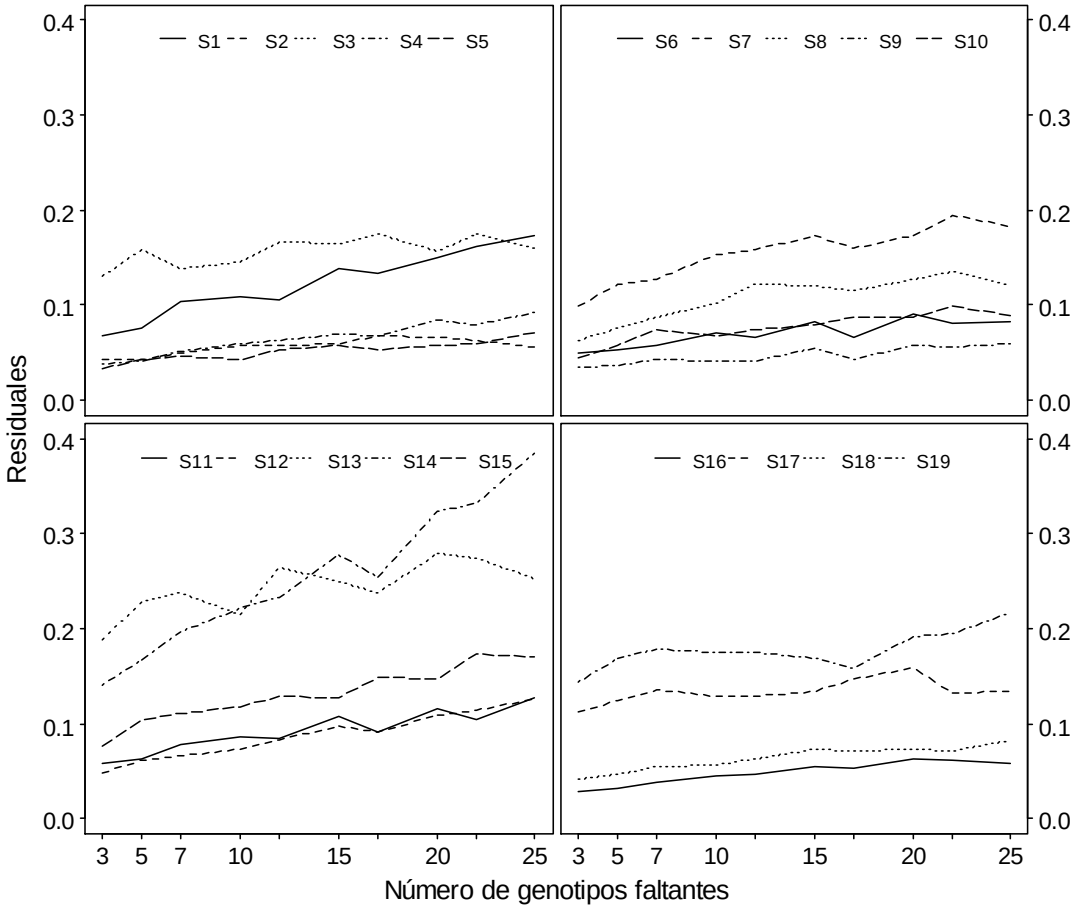


Figura 9. Valores de residuales del ajuste de Procrustes para las matrices de loadings ambiental, según número de genotipos faltantes, para los sitios S1 – S19.

2.6 DISCUSIÓN

El presente trabajo tuvo como principal finalidad el análisis de rendimiento de un ensayo multiambiente sometido a diferentes grados y tipos de pérdida aleatoria de datos, y como ésta pérdida afecta la interpretación de la interacción genotipo por ambiente.

Se pudo observar de las diversas estructuras de varianza analizadas, una caída sistemática de similaridad entre matrices cuando son sometidas a la pérdida de datos. La correlación no paramétrica del test de Mantel nos muestra como la similaridad cae con el aumento en los

datos faltantes, y que esa caída es mayor para la matriz de correlación genética que para la covarianza. A su vez, a nivel de las aleatorizaciones, en la medida que aumenta el porcentaje de pérdida lo hace también su grado de dispersión. Este comportamiento de la estructura de correlación y de la covarianza se manifestó para ambos tipos de pérdida, si bien en el caso de la pérdida de genotipos los valores son menores.

Existió una tendencia lineal en esa disminución, por lo que no fue posible trazar un punto de corte a partir del cual sería o no recomendable analizar los datos. Si bien es importante poder contestar a la pregunta de a partir de qué nivel de pérdida los resultados del análisis ya no son fiables, es necesario definir en cada situación cual es el límite a tomar, cosa que no fue abordado en el trabajo realizado.

En el caso de Procrustes, analizando el estadístico m^2 (que contempla los efectos de escala, traslación y rotación simultáneamente), vemos que aumenta en más de 6 veces su valor para la matriz de varianzas-covarianzas y en más de 4 veces para la matriz de correlación, cuando la pérdida de parcelas se incrementa de un 5% a un 50%. La variación en las matrices de loadings y scores fue importante aunque menor a la observada para la matriz de correlación. Si la pérdida es en genotipos, el estadístico m^2 se incrementa en más de 8 veces para la matriz de varianzas-covarianzas, al pasar de 3 a 25 genotipos faltantes por sitio, mientras que para las matrices de correlación, de loadings y scores el aumento es entre 3 y 4 veces su valor.

Esto sugiere que la naturaleza de la pérdida es más importante que su magnitud (importa más cuales genotipos y en que sitios se pierden). Se hace necesario entonces un estudio con otro patrón de pérdida, de tipo selectivo. En ese sentido, son importantes las afirmaciones de

Piepho y Mohring (2006) que concluyeron que la selección es ignorable en un contexto de pérdida de datos aleatorio (es decir, no necesita modelizarse) siempre que toda la información haya sido incluida en el análisis. Sus resultados revelan que es deseable utilizar la mayor cantidad posible de datos para las estimaciones de componentes de varianza, por ejemplo datos de varios años ya que las estimaciones tienen mayor precisión y la interacción cultivar por año se reduce.

Surge la pregunta sobre el efecto de datos faltantes cuando al modelo mixto de análisis con datos completos le aportamos información de parentesco. Cabe esperar que, al aumentar el número de genotipos perdidos por sitio, el efecto de la pérdida de datos sea menor dado la existencia de relaciones entre genotipos.

Por último, se analizaron los efectos principales de escala, translación y rotación de Procrustes, cuyos resultados fueron presentados fundamentalmente en los gráficos de residuales de Procrustes. De dichos gráficos surge claramente que son aquellos sitios con mayor variabilidad (los que poseen ejes de variación mayores en el biplot) los que explican en mayor medida el cambio en el ajuste de Procrustes. Parece estar asociada la disminución del factor ρ de escala con el aumento en la variabilidad residual total, de forma tal que al perder datos el ajuste se logra con factores de escala menores a la unidad.

2.7 BIBLIOGRAFÍA

Annicchiarico P. 2002. Genotype x environment interaction: challenges and opportunities for plant breeding and cultivar recommendations. FAO, Food and Agriculture Organization of the United Nations, Plant Production and Protection Paper, No. 174. Rome: FAO. 115p.

Arciniegas-Alarcón S, García-Peña M, dos Santos Dias CT, y Krzanowski WJ. 2010. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. Biometrical Letters, 47(1): 1-14.

Balzarini M. 2002. Applications of Mixed Models in Plant Breeding. En Quantitative genetics, genomics, and plant breeding, editado por Kang MS. Wallingford, UK: CABI Publishing. 353–365.

Burgueño J, Crossa J, Cotes JM, San Vicente F y Das B. 2011. Prediction assessment of linear mixed models for multienvironment trials. Crop Science, 51(3): 944-954.

Burgueño J, Crossa J, Cornelius PL, Trethowan R, Mclaren G y Krishnamachari A. 2007. Modeling additive x environment and additive x additive x environment using genetic covariances of relatives of wheat genotypes. Crop Science, 47(1): 311–320.

Crossa J, Burgueño J, Cornelius PL, Mclaren G, Trethowan R, y Krishnamachari A. 2006. Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Science, 46(4): 1722–1733.

Crossa J, Yang RC y Cornelius PL. 2004. Studying crossover genotype x environment interaction using Linear-Bilinear models and mixed models. Journal of Agricultural, Biological, and Environmental Statistics, 9(3): 362–380.

Crossa J. 1990. Statistical Analyses of Multilocation Trials. Advances in Agronomy, 44: 55–85.

Freeman GH. 1973. Statistical methods for the analysis of genotype-environment interactions. Heredity, 31(3): 339-354.

Harville DA. 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358): 320-338.

Jackson DA. 1995. PROTEST: A PROcrustean randomization TEST of community environment concordance. Ecoscience, 2(3): 297-303.

Kang MS. 2002. Genotype-Environment interaction: Progress and prospects. En Quantitative genetics, genomics, and plant breeding, editado por Kang MS. Wallingford, UK: CABI Publishing. 221–243

Kang MS y Gauch HG. 1996. Genotype -by- environment interaction. Boca Raton, FL: CRC Press. 416p.

Kelly AM, Smith AB, Eccleston JA, y Cullis BR. 2007. The accuracy of varietal selection using factor analytic models for Multi-Environment plant breeding trials. Crop Science, 47(3): 1063–1070

Manly BFJ. 2005. Multivariate statistical methods: a primer. Boca Ratón, FL: Chapman y Hall/CRC. 208p.

Peres-Neto PR, y Jackson DA. 2001. How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the mantel test. Oecologia, 129(2): 169-178.

Piepho HP y Mohring J. 2006. Selection in cultivar trials - is it ignorable? Crop Science: 46(1), 192–201.

Piepho HP. 1997. Analyzing Genotype-Environment data by mixed models with multiplicative terms. Biometrics, 53(2): 761–766.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Smith A, Cullis BR, y Thompson R. 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. The Journal of Agricultural Science, 143(06): 449–462.

Smith A, Cullis B, y Thompson R. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics, 57(4): 1138–1147

VSN International. 2010. ASReml. Hemel Hempstead, UK.

Yan W. 2013. Biplot analysis of incomplete Two-Way data. Crop Science, 53(1): 48-57.

3. DISCUSIÓN GENERAL

El análisis de datos provenientes de ensayos multiambiente tiene como característica común la presencia de cierto grado de desbalance. Al fitomejorador se le presentan varias alternativas para el análisis de sus datos, que van desde ignorar el problema (analizar sólo datos completos) hasta la aplicación de modelos complejos o estrategias para la predicción del dato faltante. En el presente trabajo se buscó una metodología que de forma directa no tuviera limitación en cuanto a la presencia de datos faltantes.

De las metodologías revisadas surge claramente el uso de un modelo mixto para el análisis. Es en el marco de dicho análisis, y por la particularidad de los datos de mejoramiento genético, donde interesa utilizar un modelo de factores analíticos para mejorar la interpretación de la interacción genotipo por ambiente, dado las implicancias fundamentales a la hora de la toma de decisiones sobre que materiales y en que ambientes se obtendrán ventajas económicas del proceso de evaluación de cultivares. De forma cotidiana el mejorador debe decidir en cuantos y en cuales ambientes utilizar, cuantas repeticiones, en cuantos años, etc., decisiones que impactan fuertemente en la economía del proceso.

De esta forma, es fundamental el poder hacer uso de toda la información que se posee, e incluso la posibilidad de no evaluar todos los materiales en todos los sitios. En la actualidad, gracias a la información genómica que se posee de los materiales probados, y la relación de parentesco entre ellos, es posible estimar la respuesta del genotipo en cualquiera de los ambientes de mejor manera.

Dado que en un programa de mejoramiento, generalmente el desbalance se genera producto de las decisiones que se toman (descarte de materiales de bajo potencial, inclusión de materiales promisorios, mantenimiento de materiales estables, etc.) es crítico poder evaluar si es más importante lo que se pierde que cuanto se pierde.

De los resultados del experimento presentado surge que en la medida que el porcentaje de datos perdidos aumenta, el resultado de nuestras estimaciones sufre cambios importantes. Evaluado a través de Procrustes, nos dice que éstas se alejan del modelo de referencia, principalmente debido a un efecto de escala. Además, importa cuál es la

estructura en la que analizamos el cambio ya que en las matrices de varianzas-covarianzas tuvo mayor efecto que en las matrices de correlaciones o en la de loadings ambientales.

Debido a que en el presente trabajo fue evaluada la pérdida aleatoria de datos no podemos afirmar lo que pasaría cuando la pérdida es dirigida o selectiva. A pesar de ello, puede decirse que cuando se eliminan materiales claves responsables de la mayor parte de la variabilidad explicada, es esperable que los patrones de respuesta cambien marcadamente. En el otro sentido, la pérdida de materiales promedio (aquellos que no se espera alta variabilidad GEI) se espera no ocasionen cambios. Por tanto, cobra importancia cuales genotipos pierdo y en que ambientes por sobre la perdida aleatoria de parcelas. En este último caso, la metodología de análisis permite estimaciones fiables en la medida en que la pérdida no supere niveles demasiado altos.

En este sentido, vale la pena como investigación a futuro, evaluar el efecto de los datos faltantes en un esquema de pérdida de tipo selectivo que se aproxime al que sucede en un esquema de prueba real. En ese contexto, generalmente el desbalance se genera por sacar materiales de bajo rendimiento, manteniendo los mejores o de comportamiento más estable. A su vez, se pueden definir diferentes escenarios: ¿es lo mismo cuando evaluamos en esquemas de selección temprana, o cuándo es un esquema avanzado? Estudios anteriores indican un efecto menor de la pérdida de datos cuando se evalúa un conjunto de datos mayor, con más genotipos y en varios años. Las posibilidades que ofrece el enfoque de modelos mixtos de incorporar información de parentesco permiten estimar la respuesta del genotipo para cualquier ambiente de mejor manera, evitando así tener que evaluar todos los materiales en todos los sitios.

Estas cuestiones, entre otras, son importantes contestar a la hora de la toma de decisiones en el marco de un programa de evaluación de materiales genéticos.

4. BIBLIOGRAFÍA

Annicchiarico P. 2002. Genotype x environment interaction: challenges and opportunities for plant breeding and cultivar recommendations. FAO, Food and Agriculture Organization of the United Nations, Plant Production and Protection Paper, No. 174. Rome: FAO. 115p.

Arciniegas-Alarcón S, García-Peña M, dos Santos Dias CT, y Krzanowski WJ. 2010. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. Biometrical Letters, 47(1): 1-14.

Balzarini M. 2002. Applications of Mixed Models in Plant Breeding. En Quantitative genetics, genomics, and plant breeding, editado por Kang MS. Wallingford, UK: CABI Publishing. 353–365.

Beeck CP, Cowling WA, Smith AB, y Cullis BR. 2010. Analysis of yield and oil from a series of canola breeding trials. part i. fitting factor analytic mixed models with pedigree information. Genome / National Research Council Canada, 53(11): 992-1001.

Bernardo R. 2010. Breeding for quantitative traits in plants, Second ed. Woodbury, MN: Stemma Press. 400p.

Burgueño J, Crossa, J, Cotes JM, San Vicente F y Das B. 2011. Prediction assessment of linear mixed models for multienvironment trials. Crop Science, 51(3): 944-954.

Burgueño J, Crossa J, Cornelius PL y Yang RC. 2008. Using factor analytic models for joining environments and genotypes without crossover genotype x environment interaction. Crop Science, 48(4): 1291–1305.

Burgueño J, Crossa J, Cornelius PL, Trethowan R., Mclaren G, y Krishnamachari A. 2007. Modeling additive x environment and additive x additive x environment using genetic covariances of relatives of wheat genotypes. Crop Science, 47(1): 311–320.

Ceretta S, y Abadie T. 2003. Avances y perspectivas del análisis de la interacción genotipo por ambiente: su contribución al estudio de la adaptación en trigo. En Estrategias y Metodologías utilizadas en el Mejoramiento de Trigo, editado por Kohli MM, Díaz M, y Castro M. Montevideo: Ed. Hemisferio Sur. 257-273.

Corbeil RR, y Searle SR. 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics, 18(1): 31–38.

Cornelius PL, Crossa J, y Seyedsadr MS. 1996. Statistical tests and estimators of multiplicative models for Genotype-by-Environment interaction. En: Genotype-by-Environment Interaction, editado por Gauch HG y Kang MS. Boca Raton, FL: CRC Press. 199-234.

Crossa J. 2012. From genotype × environment interaction to gene × environment interaction. Current genomics, 13(3): 225-244.

Crossa J, Burgueño J, Cornelius PL, Mclaren G, Trethowan R, y Krishnamachari A. 2006. Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Science, 46(4): 1722–1733.

Crossa J, Yang RC, y Cornelius PL. 2004. Studying crossover genotype x environment interaction using Linear-Bilinear models and mixed models. Journal of Agricultural, Biological, and Environmental Statistics, 9(3): 362–380.

Crossa J, y Cornelius PL. 2002. Linear-Bilinear Models for the Analysis of Genotype-Environment Interaction. En Quantitative genetics, genomics, and plant breeding, editado por Kang MS. Wallingford, UK: CABI Publishing. 305–321.

Crossa J. 1990. Statistical Analyses of Multilocation Trials. Advances in Agronomy, 44: 55–85.

Cullis BR, Smith AB, Beeck CP, y Cowling WA. 2010. Analysis of yield and oil from a series of canola breeding trials. part II. exploring variety by environment interaction using factor analysis. Genome / National Research Council Canada, 53(11): 1002–1016.

Dietz EJ. 1983. Permutation tests for association between two distance matrices. Systematic Zoology 32(1): 21-26.

Ferreira DF, Demétrio CG, Manly BF, de Almeida Machado A, y Vencovsky R. 2006. Statistical models in agriculture: biometrical methods for evaluating phenotypic stability in plant breeding. Cerne, 12(4): 373-388.

Finlay KW y Wilkinson GN. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research, 14, 742–754.

Flores F, Moreno MT, y Cubero JI. 1998. A comparison of univariate and multivariate methods to analyze GxE interaction. Field Crops Research, 56(3): 271–286.

Freeman GH. 1973. Statistical methods for the analysis of genotype-environment interactions. Heredity, 31(3): 339–354.

Gabriel KR. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58(3): 453–467.

Gauch HG, Piepho HP y Annicchiarico P. 2008. Statistical analysis of yield trials by AMMI and GGE: Further considerations. Crop Science, 48(3): 866–889

Gauch HG, y Zobel RW. 1996. Ammi analysis of yield trials. En Genotype -by- environment interaction, editado por Kang MS y Gauch HG, 85-122. Boca Raton, FL: CRC Press, 85-122.

Gauch HG, y Zobel RW. 1990. Imputing missing yield trial data. Theoretical and Applied Genetics, 79(6): 753–761.

Gauch HG. 1988. Model selection and validation for yield trials with interaction. Biometrics, 44(3): 705–715.

Gollob H. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika, 33(1): 73-115.

Gower J. 1975. Generalized procrustes analysis. Psychometrika, 40(1): 33–51.

Hanson WD. 1970. Genotypic stability. Theoretical and Applied Genetics, 40(5): 226–231.

Hartley HO, y Hocking RR. 1971. The analysis of incomplete data. Biometrics, 27(4): 783–823.

Harville DA. 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358): 320–338.

Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31(2): 423-447.

Hill J, Becker HC, y Tigerstedt PMA. 1998. Genotype—environment interactions: analysis and problems. En Quantitative and Ecological Aspects of Plant Breeding, Plant Breeding, London: Chapman y Hall. 155-186.

Hill J. 1975. Genotype-environment interaction - a challenge for plant breeding. The Journal of Agricultural Science, 85(3): 477–493.

Jackson DA. 1995. PROTEST: A PROcrustean randomization TEST of community environment concordance. Ecoscience 2(3):297-303.

Jennrich RI, y Schluchter MD. 1986. Unbalanced Repeated-Measures models with structured covariance matrices. Biometrics, 42(4): 805–820.

Kang MS. 2004. Breeding: Genotype-by-Environment interaction. En: Encyclopedia of Plant and Crop Science. Nueva York: Marcel Dekker Inc. 218–221.

Kang MS, Balzarini MG, y Guerra JL. 2004. Genotype-by-Environment interaction. En Genetic Analysis of Complex Traits Using SAS, editado por Saxton A. Cary, NC: SAS Institute Inc. 69-96.

Kang MS. 2002. Genotype-Environment interaction: Progress and prospects. En Quantitative genetics, genomics, and plant breeding, editado por Kang MS. Wallingford, UK: CABI Publishing. 221–243

Kang MS. 1997. Using Genotype-by-Environment Interaction for Crop Cultivar Development. Advances in Agronomy, 62: 199-252.

Kang MS y Gauch HG. 1996. Genotype -by- environment interaction. Boca Raton, FL: CRC Press. 416p.

Kelly AM, Cullis BR, Gilmour AR, Eccleston JA, y Thompson R. 2009. Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genetics Selection Evolution, 41(1): 33-41.

Kelly AM, Smith AB, Eccleston JA, y Cullis BR. 2007. The accuracy of varietal selection using factor analytic models for Multi-Environment plant breeding trials. Crop Science, 47(3): 1063–1070.

Krzanowski WJ. 1987. Selection of variables to preserve multivariate data structure, using principal components. Applied Statistics, 36(1): 22-33.

Littell RC. 2002. Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. Journal of Agricultural, Biological, and Environmental Statistics, 7(4): 472–490.

Manly BFJ. 2005. Multivariate statistical methods: a primer. Boca Ratón, FL: Chapman y Hall/CRC. 208p.

Mariotti JA. 1994. La interacción genotipo-ambiente, su significado e importancia en el mejoramiento genético y en la evaluación de cultivares. Serie monográfica n° 1. INTA-CRTS. 38p.

McLean RA, Sanders WL, y Stroup WW. 1991. A unified approach to mixed linear models. The American Statistician, 45(1): 54–64.

Meyer K. 2009. Factor-analytic models for genotype x environment type problems and structured covariance matrices. Genetics Selection Evolution, 41(1): 21-32.

Patterson HD y Thompson R. 1971. Recovery of Inter-Block information when block sizes are unequal. Biometrika, 58 (3): 545-554.

Peres-Neto PR, y Jackson DA. 2001. How well do multivariate data sets match: the advantages of a procrustean superimposition approach over the mantel test. Oecologia 129(2): 169-178.

Piepho HP y Mohring J. 2006. Selection in cultivar trials - is it ignorable? Crop Science, 46(1): 192–201.

Piepho HP. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics, 97(1): 195–201.

Piepho HP. 1997. Analyzing Genotype-Environment data by mixed models with multiplicative terms. Biometrics, 53(2): 761–766.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rohlf FJ, y Slice D. 1990. Extensions of the procrustes method for the optimal superimposition of landmarks. Systematic Zoology 39(1): 40-59.

Roozeboom KL, Schapaugh WT, Tuinstra MR, Vanderlip RL, y Milliken GA. 2008. Testing wheat in variable environments: Genotype, environment, interaction effects, and grouping test locations. Crop Science, 48(1): 317–330.

Rubin DB. 1976. Inference and missing data. Biometrika, 63 (3): 581-592.

Schneider JW y Borlund P. 2007. Matrix comparison, part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. Journal of the American Society for Information Science and Technology. 58(11): 1596-1609.

Shukla GK. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. Heredity, 29(2): 237–245.

Smith A, Cullis B, y Thompson R. 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. The Journal of Agricultural Science, 143(6): 449–462.

Smith A, Cullis B, y Thompson R. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics, 57(4): 1138–1147.

van Eeuwijk FA. 1995. Linear and bilinear models for the analysis of multi-environment trials: I. an inventory of models. Euphytica, 84(1): 1–7.

VSN International. 2010. ASReml. Hemel Hempstead, UK.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, y Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using procrustes analysis. Statistical applications in genetics and molecular biology 9(1): Article 13.

Xu Y. 2010. Genotype-by-environment interaction. Cambridge, MA: CAB International. 318-416.

Yan W. 2013. Biplot analysis of incomplete Two-Way data. Crop Science, 53(1): 48-57.

Yan W, Kang MS, Ma B, Woods S, y Cornelius PL. 2007. GGE biplot vs. AMMI analysis of Genotype-by-Environment data. Crop Science, 47(2): 643-653.

Yan W, y Tinker NA. 2005. An integrated biplot analysis system for displaying, interpreting, and exploring genotype × environment interaction. Crop Science, 45(3): 1004-1016.

Yang RC, Crossa J, Cornelius PL, y Burgueño J. 2009. Biplot analysis of genotype-environment interaction: Proceed with caution. Crop Science, 49(5): 1564–1576.

Yang RC. 2007. Mixed-Model analysis of crossover Genotype-Environment interactions. Crop Science, 47(3): 1051–1062.

Yang RC. 2002. Likelihood-Based analysis of Genotype-Environment interactions. Crop Science, 42(5): 1434–1440.

Zobel RW, Wright MJ, y Gauch HG. 1988. Statistical analysis of a yield trial. Agronomy Journal, 80(3): 388-393.

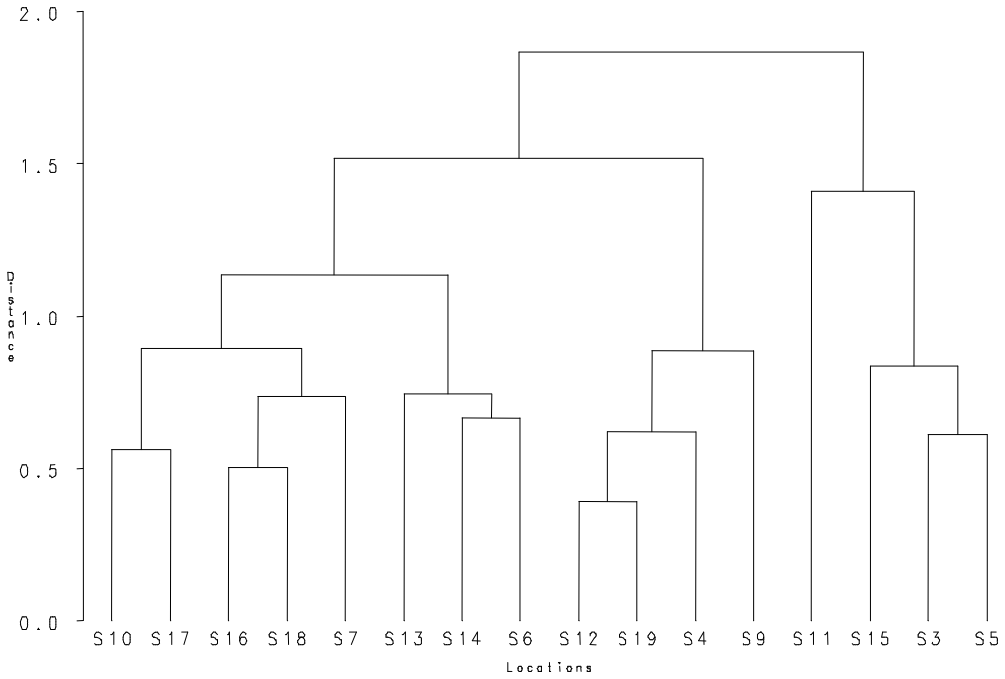
5. ANEXOS

Tabla de sitios y genotipos

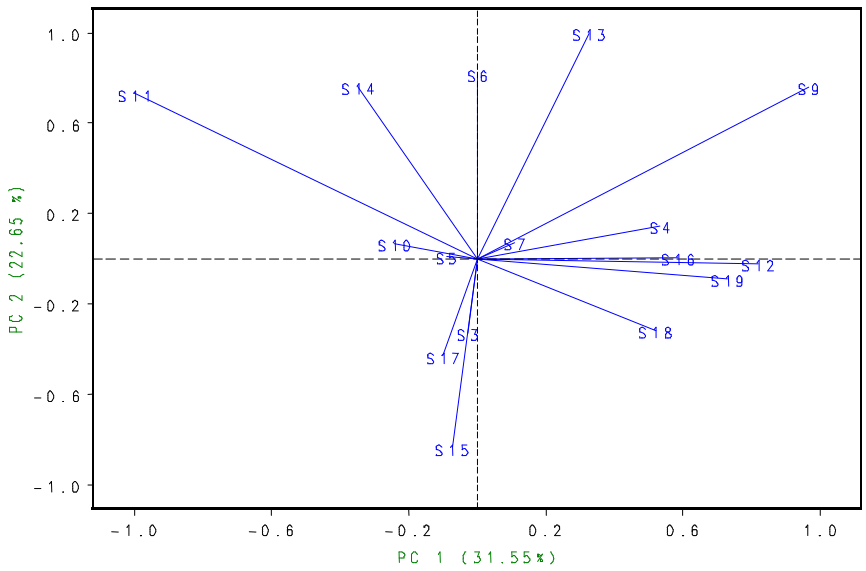
SITIO	CODIGO	DESCRIPCION	PAIS
1	10022	MOREDOU	SOUTH AFRICA
2	22501	NWRP- BHAIRAHWA	NEPAL
3	22611	WHEAT RESEARCH ITUTE	PAKISTAN
4	22612	BARANI	PAKISTAN
5	22614	DERA ISMAIL KHAN	PAKISTAN
6	27101	SAN-PA-TONG	THAILAND
7	27104	SAMOENG UPLAND RICE AND TEMP. CEREALS	THAILAND
8	27118	PANG MA PHA	THAILAND
9	29101	KAZAKH GRAIN	KAZAKHSTAN
10	42301	MIXTECA OAXAQUENA	MEXICO
11	45403	ESCUELA AGR. PANAMERICANA-EL ZAMORANO	HONDURAS
12	51004	MARCOS JUAREZ	ARGENTINA
13	51205	P. UNIV. CATOLICA DE CHILE	CHILE
14	53002	SAN BENITO	BOLIVIA
15	53014	CHUQUISACA	BOLIVIA
16	62405	KHARKOV	UKRAINE
17	65001	KENTZIKO THERMI	GREECE
18	65301	PBS ALENTEJO	PORTUGAL
19	65434	TORREGROSSA/BELLOC	SPAIN

GENOTIPO	IDENTIFICACION
1	SITTA
2	DHARWAR DRY
3	BAVIACORA M 92
4	NESSER
5	TUI
6	F60314.76/MRL//CNO79
7	FIRETAIL
8	PFAU/BOW//VEE#9
9	PASTOR
10	PFAU/VEE#5
11	ATTILA
12	KEA/BUC//FCT
13	PFAU/VEE#9
14	BOW//BUC/BUL
15	BAU/OPATA
16	PRINIA
17	BABAX
18	MRL/BUC//VEE#7
19	CHIL//ALD/PVN
20	PSN/BOW//SERI
21	FINK/BUC
22	CHIL/BUC
23	CHIL/BUC
24	IL-75-2264/4/CAR//KAL/BB/3/NAC/5/GAA
25	PIK/OPATA
26	BJY/COC//PRL/BOW
27	OPATA/KILL
28	JUN/BOMB
29	MIMUS
30	PASTOR/OPATA
31	SITTA*2//PSN/BOW
32	GEN/3/GOV/AZ//MUS/4/BUC/MOR/5/HD2359/3/GOV/AZ//MUS
33	VEE#5/SARA//OPATA/3/OPATA/BOW
34	OPATA/BOW//BAU/3/OPATA/BOW
35	OPATA/BOW*2//BUC/MOR
36	PASTOR*2/OPATA
37	PASTOR*2/OPATA
38	TAM200/TRAP#1
39	TJB368.251/BUC//CUPE
40	TJB368.251/BUC//BUC/CHRC
41	RL6043/4*NAC
42	TIA.2
43	GEN*3/PVN
44	MYNA/VUL//JUN
45	IRENA
46	CHIL//ALD/PVN
47	URES/JUN//KAUZ
48	URES/PRL
49	ND/VG9144//KAL/BB/3/YACO/4/CHIL

Dendrograma jerárquico y PCA *Biplot*, utilizando como medida de distancia la diferencia entre matriz de identidad y matriz de correlaciones genéticas entre sitios† (calculada de Cooper *et al*., 1996).



PCA Biplot

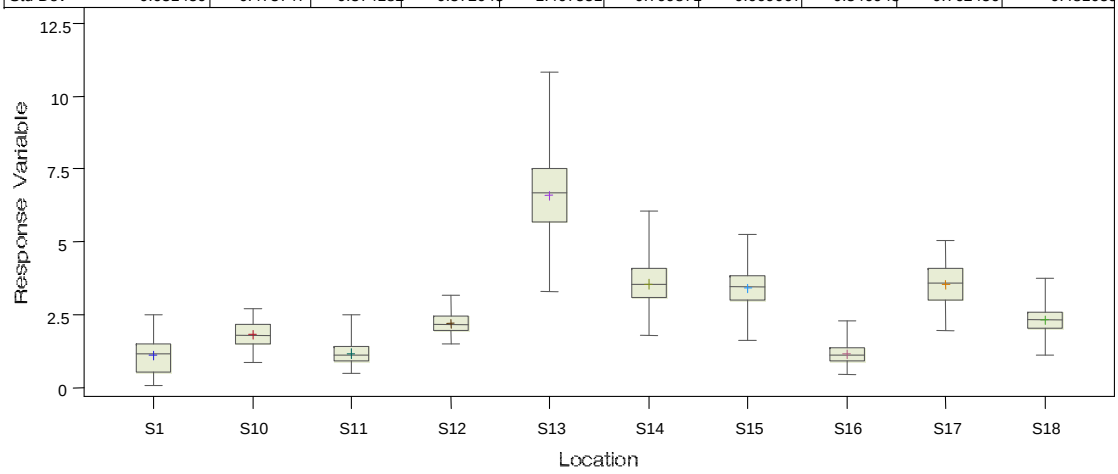


† *No incluye sitios con heredabilidad menor a 0.05*

Gráficos de *boxplot* mostrando la distribución de las observaciones para la variable rendimiento (t/ha) según los 19 sitios del análisis. Para cada sitio se brinda el indicador Mínimo (Min), Media (Mean), Máximo (Max) y Desvío Estándar (Std Dev).

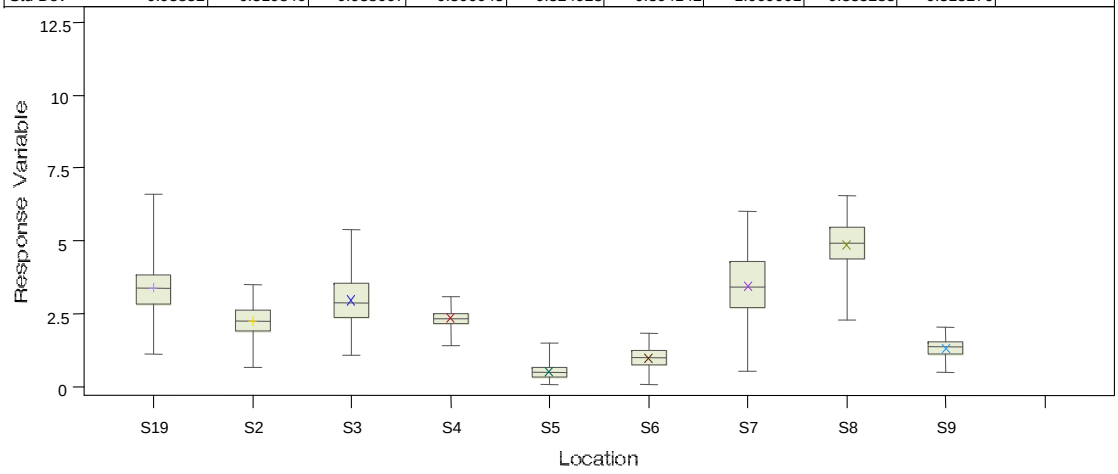
Basic Statistics by Location

	S1	S10	S11	S12	S13	S14	S15	S16	S17	S18
Min	0.064813	0.8571	0.48343	1.49985	3.29165	1.77236	1.61	0.433329	1.96146	1.1112
Mean	1.094688	1.807899	1.163362	2.200971	6.585432	3.552835	3.404796	1.142846	3.532629	2.310276
Max	2.47678	2.71415	2.48883	3.14968	10.8333	6.03937	5.256	2.28887	5.03826	3.76419
Std Dev	0.631459	0.473747	0.374182	0.372043	1.467551	0.799871	0.669067	0.346943	0.762456	0.481059



Basic Statistics by Location

	S19	S2	S3	S4	S5	S6	S7	S8	S9
Min	1.12231	0.672	1.06656	1.41695	0.08335	0.08	0.515	2.283	0.50001
Mean	3.392902	2.251592	2.96637	2.357954	0.520512	0.971694	3.43699	4.877495	1.328744
Max	6.58108	3.488	5.36613	3.08395	1.5003	1.84	6.01	6.574	2.04766
Std Dev	0.93852	0.519846	0.938667	0.306045	0.324928	0.394142	1.069001	0.855283	0.315276



Gráficos de biplot del modelo FA(2) incluyendo 19 sitios (S1-S19) y 49 genotipos (1-49) con pérdida aleatoria de parcelas de: a) 10%, b) 25%. c) 40% y d) 50%.