GOBIERNO DE CHILE
FUNDACIÓN PARA LA
INNOVACIÓN AGRARIA

INFORME TÉCNICO Y DE DIFUSIÓN

Curso
INTRODUCCION AL ANALISIS MULTIVARIADO

DR. JOSE CROSSA
JEFE UNIDAD DE BIOMETRIA Y ESTADISTICA
CENTRO INTERNACIONAL DE MAIZ Y TRIGO (CIMMYT)
MEXICO

12 al 14 de octubre de 2004

FACULTAD DE CIENCIAS AGRONOMICAS, UNIVERSIDAD DE CHILE

GOBIERNO DE CHILE
FUNDACION PARA LA
INNOVACIÓN AGRARIA

INFORME TÉCNICO Y DE DIFUSIÓN

Fecha de entrega del Informe
Nombre del coordinador de la ejecución
EDMUNDO ACEVEDO HINOJOSA
Firma del Coordinador de la Ejecución


1. ANTECEDENTES GENERALES DE LA PROPUESTA
Nombre de la propuesta
INTRODUCCION AL ANALISIS MULTIVARIADO
Código
FIA-FR-V-2004-1-A-010
Postulante o Postulantes
Facultad de Ciencias Agronómicas, Universidad de Chile
Entidad Patrocinante o Responsable
Facultad de Ciencias Agronómicas, Universidad de Chile
Lugar de Formación (País, Región, Ciudad, Localidad)
Chile, Región Metropolitana, Santiago, La Pintana
Tipo o Modalidad de Formación (curso, pasantía, seminario, entre otros)
Curso
Fecha de realización (Inicio y término)
12 al 14 de octubre de 2004

2. ALCANCES Y LOGROS DE LA PROPUESTA

Entrenar a un grupo de profesionales del agro que se desempeñen en agencias públicas y privadas de docencia, investigación y extensión en el uso de las nuevas técnicas de análisis estadístico multivariado.

Objetivos Específicos:

- Contrastar diseños de experimentos repetidos y no repetidos.
- Realizar análisis espaciales y de interacción GXE.
- Conocer y aplicar modelos lineales-bilineales para estudio de estabilidad.
- Conocer la estructura y uso de biplots y sus aplicaciones.
- Estudiar los modelos SHMM y SREG para agrupar ambientes y genotipos sin interacción.
- Estudiar la GXE de entrecruzamiento.

Se reunió un grupo diverso de profesionales del agro que trabajan en instituciones de Investigación y Extensión Públicas y Privadas (U. de Chile, U. de Valdivia, INIA, Semillas Pioneer Chile Ltda., Semillas CIS), junto a estudiantes de pregrado y postgrado. Estas personas trabajan en el área del mejoramiento genético y selección de variedades, por lo que los conocimientos adquiridos durante el curso serán de gran utilidad en la mejor comprensión de la información obtenida de sus ensayos.

Los Objetivos Específicos planteados al inicio del curso, que correspondían a las materias a tratar, se cumplieron a cabalidad.

Usar de manera más eficiente de la información de ensayos de experimentación agrícola, ya sean de investigación o de extensión. Esto permitirá un avance más rápido y preciso de la investigación agronómica. Los participantes se invitarán con el criterio de que a través de su labor difundan estas técnicas con rapidez en el país.

Los asistentes se introdujeron en las técnicas de análisis estadístico de punta, las que se perfeccionarán mediante la aplicación práctica en los distintos niveles de trabajo de cada uno.

La diversidad de origen de los asistentes asegura el uso de estos conocimientos a nivel de docencia e investigación.

Establecimiento de un vinculo permanente entre el Dr. José Crossa y los participantes, de manera de interactuar con él frente a dudas, problemas de aplicación o interpretación, referentes al curso o a otras materias estadísticas.

(3)

Los asistentes cuentan con una herramienta más de análisis que les facilitará la interpretación de la información de campo, en particular la interacción Genotipo x Medioambiente. A futuro les permitirá hacer diseños de campo más eficientes en la estimación del efecto ambiental.

(4)

El curso fue una introducción a técnicas avanzadas de análisis estadístico, y en la medida que se empiecen a utilizar sería necesario una profundización de los conceptos que se entregaron, lo que justificaría una nueva visita del Dr. José Crossa.

3. ASPECTOS RELACIONADOS CON LA ORGANIZACIÓN Y EJECUCIÓN DE LA PROPUESTA

Fecha	Actividad	Objetivo	Lugar
12/10/2004	Diseño de experimentos repetidos Bloques completos Bloques incompletos: - latice tradicionales - alpha latices - hileras-columna - latinizados Ejemplo del uso de latice en ensayos de maiz en condiciones de sequía		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
12/10/2004	Diseño de experimento no repetidos ¿por que? cuando y ¿como?		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
12/10/2004	Análisis espacial Modelo autoregresivo en el sentido de las hileras y las columnas Splines y otros		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
13/10/2004	Interaccion GxE		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
13/10/2004	Modelo de regresión simple para estudiar la estabilidad de genotipos		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
13/10/2004	Modelo lineales-bilineales para estudiar la estabilidad de genotipos		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
13 y 14 10/2004	Biplots de los modelos lineales-bilineales		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
14/10/2004	Los modelos SHMM y SREG para agrupar ambientes y genotipos sin interacción		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile
14/10/2004	GxE de entrecruzamiento		Area de Computación Facultad de Ciencias Agronómicas, Universidad de Chile

Copiar toda esta columna

Institución/ Empresa/Organi-zación	Persona de Contacto	Cargo	Fono/Fax	Dirección	E-mail
Universidad Austral	Daniel Calderini	Profesor	(63) 221723 Fax: (63) 221233	Instituto de Producción y Sanidad Vegetal, Campus Isla Teja, Valdivia. X Región	danielcalderini@uach.cl
INIA	Gabriel Saavedra	Coordinador e Investigador del Depto. de Hortalizas y Cultivos	02-7575154 02-5417667	Santa Rosa 11610, La Pintana	gsaavedr@platina.inia.cl
Semillas Pioneer Chile Ltda.	Gastón Delard		3622337		gaston.delard@pioneer.com
Semillas CIS Ltda..	Rodrigo Carvacho	Jefe Programa Investigación Remolacha	2035193	Av El Bosque Norte 0440, piso 8	rcarvacho@císchile.cl

GOBIERNO DE CHILE
FUNDACIÓN PARA LA INNOVACIÓN AGRARIA

Material Recopilado

Junto con el informe técnico se debe entregar un set de todo el material recopilado durante la actividad de formación (escrito y audiovisual) ordenado de acuerdo al cuadro que se presenta a continuación (deben señalarse aquí las fotografías incorporadas en el punto 4):

Tipo de Material	Nº Correlativo (si es necesario)	Caracterización (título)
Artículo	1	Crossa, J., Cornelius, P.L. and Weikai Yan. 2002. Biplots of Linear-Bilinear Models for studying crossover genotype environment interaction. Crop Science 42:619–633.
	2	Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramirez, M. and Talbot, M. 1998. Interpreting Genotype x Environment interaction in wheat by Partial Least Squares regression. Crop Science 38:679–689.
	3	*Linear-Bilinear models for the análisis of Genotype-Environment interaction. 2002. Crossa, J. and Cornelius, P.L. In: Quantitative Genetics, genomics and Plant Breeding. M.S Kang (Ed).*
	4	Trethowan, R., Crossa, J., Van Ginkel, M. and Rajaram, S. 2001. Relationships among Bread Wheat International Yield Testing Locations in Dry Areas. Crop Sci. 41:1461–1469.
	5	Yan, W., Cornelius, P., Crossa, J. and L. A. Hunt. 2001. Two Types of GGE Biplots for analyzing Multi-Environment trial data. Crop Sci. 41:656–663.
	6	Vargas, M., Crossa, J., Van Eeuwijk, F., Ramirez, M. and Sayre, K. 1999. Using partial Least Squares Regression, Factorial Regression, and AMMI Models for interpreting Genotype x Environment interaction. Crop Sci. 39:955–967.
	7	Crossa, J., Van Eeuwijk, F., Cornelius, P. and Vargas, M. Linear, Bilinear rmodels for analizing Genotipe x Environment interaction.
Foto		
Libro (manual)		User's guide for spatial analysis of field variety trials using ASREML. A. Cadena, J. Burgueño, J. Crossa

Diapositiva		
CD		

4. PROGRAMA DE DIFUSIÓN EJECUTADO

Programa de difusión ejecutado

En esta sección se deberán describir detalladamente las actividades de difusión realizadas, tales como publicaciones, charlas, seminarios u otras actividades similares, comparando con el programa establecido inicialmente en la propuesta. Se deberá también describir y adjuntar el material de difusión preparado y/o distribuido en dichas actividades.

La información a entregar sobre cada actividad de difusión es la siguiente:

- Tipo de actividad realizada y objetivo principal (incluye elaboración de publicaciones)
- Fecha y lugar de realización
- Temas tratados o exposiciones realizadas
- Destinatarios de la actividad: especificar el tipo y número de personas que asistieron a la actividad (productores, académicos, investigadores, profesionales, técnicos, etc.). Se deberá adjuntar el listado de asistentes según formato indicado más adelante.
- Nombre y tipo de las organizaciones u otras instituciones relevantes en el tema o sector que tuvieron representación en la asistencia al evento.
- Identificación de los expositores que estuvieron a cargo de las presentaciones, indicando su vinculación con la iniciativa y lugar de trabajo
- Indicar si se trató de una actividad abierta a todos los interesados, abierta a quienes se inscribieron previamente, o limitada a quienes fueron especificamente invitados.
- En el caso de los seminarios, deberá adjuntarse el Programa de la actividad que se realizó.

Actividad realizada:
Curso

Fecha:
12 al 14 de Octubre de 2004

Temas tratados:
Diseño estadistico, Interacción Genotipo x Medioambiente, Análisis espacial, Análisis de estabilidad de rendimiento, Análisis de componentes principales, Biplots (AMMI, SREG).

Asistentes:
Participaron de este curso 24 personas, entre los que se cuentan profesionales que trabajan en Investigación, Mejoramiento Genético y Selección de variedades. Estos pertenecen a distintas instituciones, publicas y privadas. Además participaron estudiantes de pregrado de la carrera de Ingeniería Agronómica de la Universidad de Chile, estudiantes de postgrado del programa de Doctorado en Ciencias Silvoagropecuarias de la U. de Chile y del Programa de

Magister de la U. Austral

Instituciones asistentes:
Instituciones Publicas: Facultad de Ciencias Agronómicas de la Universidad de Chile, Facultad de Ciencias Forestales de la Universidad de Chile, Facultad de Ciencias Agrarias de la Universidad Austral, Instituto Nacional de Investigaciones Agrícolas (INIA).
Instituciones Privadas: Semillas Pioneer Chile Ltda. y Semillas CIS.

Expositor:
Dr. José Crossa, Jefe de la Unidad de Biometría y Estadística del Centro Internacional de Mejoramiento del Maíz y Trigo (CIMMYT), México.

Tipo de Invitación:
Las invitaciones fueron dirigidas, de manera de reunir a profesionales y académicos que se enfrenten a la problemática de la interpretación de la interacción Genotipo x Medioambiente.

Material entregado en las actividades de difusión

Entregar un listado del material elaborado y distribuido con motivo de la actividad o material audiovisual exhibido como video, datashow, entre otros.

Además, se debe entregar adjunto al informe un set de todo el material entregado en las actividades de difusión (escrito y audiovisual) ordenado de acuerdo al cuadro que se presenta a continuación.

También se deben adjuntar fotografías correspondientes a la actividad desarrollada. El material se debe adjuntar en forma impresa y en un medio magnético (disquet o disco compacto).

Tipo de material	Nombre o identificación	Preparado por	Cantidad
Disco compacto	Clases dictadas por el Dr. Crossa	José Crossa	1
Disco compacto	Ejemplos de conceptos entregados en clases, programa estadístico ASREML (versión libre), ejercicios prácticos de aplicación de conceptos. Artículos relacionados y manual de uso del programa estadístico escrito por el Dr. Crossa.	José Crossa	1

Participantes en actividades de difusión Es necesario registrar los antecedentes de todos los asistentes que participaron en las actividades de difusión. El listado de asistentes a cualquier actividad deberá al menos contener la siguiente información:	
Nombre	Edmundo
Apellido Paterno	Acevedo
Apellido Materno	Hinojosa
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11315, La Pintana. Reg. Metropolitana
Fono y Fax	(2) 678 5858
E-mail	eacevedo@uchile.cl edmundoacevedo@vtr.net
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Profesor titular
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
2	
Nombre	Paola
Apellido Paterno	Silva
Apellido Materno	Candia
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11315, La Pintana. Reg. Metropolitana
Fono y Fax	(2) 678 5858
E-mail	psilva@uchile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas

RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Profesor asistente
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
3	
Nombre	Herman
Apellido Paterno	Silva
Apellido Materno	Robledo
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11315, La Pintana. Reg. Metropolitana
Fono y Fax	(2) 678 5858
E-mail	hsilva@uchile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Profesor asociado
Rubro, área o sector a la cual se vincula o en la que trabaja	Ecofisiológia vegetal
4	
Nombre	Alberto
Apellido Paterno	Mansilla
Apellido Materno	Martínez
RUT Personal	
Dirección, Comuna y Región	Luis Durand 4066. Santiago, RM
Fono y Fax	221 3961 678 5802
E-mail	amansill@uchile.cl
Nombre de la organización, empresa o	Universidad de Chile. Fac. Ciencias Agronómicas

institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Profesor Titular de Estadística
Rubro, área o sector a la cual se vincula o en la que trabaja	Modelos Biológicos. Biometría
5	
Nombre	Eduardo Enrique
Apellido Paterno	Martínez
Apellido Materno	Herrera
RUT Personal	
Dirección, Comuna y Región	Pje. Cosmos 2367, Conchalí, Santiago
Fono y Fax	7352296, 6785858
E-mail	emartine@uchile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante Programa de Doctorado en Ciencias Silvoagropecuarias y Veterinarias de la U. De Chile
Rubro, área o sector a la cual se vincula o en la que trabaja	Rotaciones de Cultivos
6	
Nombre	Susana Rebeca
Apellido Paterno	Valle
Apellido Materno	Toledo
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11.315 La Pintana

Fono y Fax	6785858
E-mail	susanavallet@yahoo.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante – Tesista
Rubro, área o sector a la cual se vincula o en la que trabaja	Trabajo de Tesis de Ingeniero Agrónomo
7	
Nombre	Mauricio Felipe
Apellido Paterno	Ortiz
Apellido Materno	Lizana
RUT Personal	
Dirección, Comuna y Región	Molina 668, Buin. Región Metropolitana
Fono y Fax	(2) 678 5858
E-mail	mauricioortiz@chilesat.net
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Ingeniero Agrónomo
Rubro, área o sector a la cual se vincula o en la que trabaja	Resistencia a estrés hídrico

8	
Nombre	Marcela
Apellido Paterno	Opazo
Apellido Materno	Illanes
RUT Personal	
Dirección, Comuna y Región	Venezuela 8814, La Florida, Santiago
Fono y Fax	(2) 6785858
E-mail	maoi_opazo@123mail.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	6
Cargo o actividad que desarrolla	Ingeniero Agrónomo
Rubro, área o sector a la cual se vincula o en la que trabaja	Producción de semillas
9	
Nombre	Juan Manuel
Apellido Paterno	Barrios
Apellido Materno	Martinez
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11.315 La Pintana
Fono y Fax	6785857, 6785905
E-mail	jbarrios@uchile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Fac. Ciencias Forestales
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	60.910.000-1

Cargo o actividad que desarrolla	Profesor adjunto
Rubro, área o sector a la cual se vincula o en la que trabaja	Modelos de la Investigación Operativa. Computación
10	
Nombre	Javiera
Apellido Paterno	González
Apellido Materno	Cruz
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11315, La Pintana, Región Metropolitana
Fono y Fax	(2) 6785867
E-mail	javigonz@uchile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Facultad de Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante Programa de Doctorado en Ciencias Silvoagropecuarias y Veterinarias de la U. De Chile
Rubro, área o sector a la cual se vincula o en la que trabaja	Fisiología de cultivos
11	
Nombre	Carolina Ivon
Apellido Paterno	Rivera
Apellido Materno	Montoya
RUT Personal	
Dirección, Comuna y Región	Algarrobo 7978, La Granja, Santiago
Fono y Fax	5254367 – 97291014
E-mail	criveramont@yahoo.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Fac. Ciencias Agronómicas Universidad de Chile

RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante Ingeniería Agronómica
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
12	
Nombre	María Verónica
Apellido Paterno	Muñoz
Apellido Materno	Muñoz
RUT Personal	
Dirección, Comuna y Región	Pasaje 50 casa 2037, Conchalí
Fono y Fax	08-5444603
E-mail	veroagr@hotmail.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad de Chile, Fac. Ciencias Agronómicas
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Tesista de Ingeniería Agronómica
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
13	
Nombre	Carlo César
Apellido Paterno	Montes
Apellido Materno	Verdugo
RUT Personal	
Dirección, Comuna y Región	Salar de Llamara #10263 La Florida
Fono y Fax	2827691 / 08 5172271
E-mail	Carlomontes@hotmail.com
Nombre de la organización, empresa o	Universidad de Chile

institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante Ingeniería Agronómica
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
14	
Nombre	Daniel F.
Apellido Paterno	Calderini
Apellido Materno	Rosso
RUT Personal	
Dirección, Comuna y Región	Instituto de Producción y Sanidad Vegetal, Campus Isla Teja, Valdivia. X Región
Fono y Fax	(63) 22-1723 Fax: (63) 22-1233
E-mail	Danielcalderini@uach.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Facultad de Ciencias Agrarias, Universidad Austral de Chile
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Profesor
Rubro, área o sector a la cual se vincula o en la que trabaja	Fisiología de cultivos anuales
15	
Nombre	Patricio Alejandro
Apellido Paterno	Sandaña
Apellido Materno	Gómez
RUT Personal	
Dirección, Comuna y Región	Colon 577 Coyhaique

Fono y Fax	(63) 211864 (67) 235774
E-mail	patriciosandana@uach.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad Austral de Chile
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante de Magíster en Ciencias Vegetales
Rubro, área o sector a la cual se vincula o en la que trabaja	Agronomía
16	
Nombre	Claudia Isabel
Apellido Paterno	Harcha
Apellido Materno	Cortés
RUT Personal	
Dirección, Comuna y Región	Campus Isla Teja s/n
Fono y Fax	Secretaria: (63) 221232 Fax: (63) 221233 Celular: 0-94515395
E-mail	claudiaharcha@uach.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Universidad Austral de Chile
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante de Magister en Ciencias Vegetales Mención Fisiología Vegetal
Rubro, área o sector a la cual se vincula o en la que trabaja	Agricultura Investigación
17	
Nombre	Erika Roxana

Apellido Paterno	Salazar
Apellido Materno	Suazo
RUT Personal	
Dirección, Comuna y Región	Sta Rosa 11610, La Pintana, Santiago
Fono y Fax	56 2 7575204 56 2 5416687(FAX)
E-mail	esalazar@platina.inia.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Instituto De Investigaciones Agropecuarias Cri-La Platina
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Investigador, Encargada Banco Activo de Germoplasma
Rubro, área o sector a la cual se vincula o en la que trabaja	Recursos Genéticos
18	
Nombre	Gabriel
Apellido Paterno	Saavedra
Apellido Materno	Del Real
RUT Personal	
Dirección, Comuna y Región	Santa Rosa 11610 – La Pintana
Fono y Fax	02-7575154 y 02-5417667
E-mail	gsaavedr@platina.inia.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Instituto de Investigaciones Agropecuarias – CRI La Platina
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Coordinador e Investigador del Depto. de Hortalizas y Cultivos
Rubro, área o sector a la cual se vincula o	Mejoramiento genético de maíz choclero y tomate

en la que trabaja	para procesamiento.
19	
Nombre	Juan Eduardo
Apellido Paterno	Zarhi
Apellido Materno	Salim-Hanna
RUT Personal	
Dirección, Comuna y Región	Coyancura 2241 piso 3, Providencia
Fono y Fax	3622352 / 8249890
E-mail	juan.zarhi@pioneer.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Semillas Pioneer Chile Ltda..
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agricola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Agrónomo
Rubro, área o sector a la cual se vincula o en la que trabaja	Investigación y Producción de Semillas
20	
Nombre	Andrea Marisol
Apellido Paterno	Salinas
Apellido Materno	Rubio
RUT Personal	
Dirección, Comuna y Región	Santa Filomena , Buin
Fono y Fax	3622384
E-mail	andrea.salinas@pioneer.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Semillas Pioneer Chile Ltda..
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	

Cargo o actividad que desarrolla	Agrónomo
Rubro, área o sector a la cual se vincula o en la que trabaja	Investigación y Producción de Semillas
21	
Nombre	Gaston
Apellido Paterno	Delard
Apellido Materno	Rodriguez
RUT Personal	
Dirección, Comuna y Región	Parcela 2fd Camino La Esperanza, Pirque
Fono y Fax	3622337
E-mail	gaston.delard@pioneer.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Semillas Pioneer Chile Ltda.
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Investigador asociado
Rubro, área o sector a la cual se vincula o en la que trabaja	Investigación en maíz
22	
Nombre	Miguel
Apellido Paterno	Ibáñez
Apellido Materno	Vial
RUT Personal	
Dirección, Comuna y Región	Camino Las Flores 11.929, Las Condes
Fono y Fax	2141701
E-mail	miguel.ibanez@pioneer.com
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Semillas Pioneer Chile LTDA.
RUT de la organización, empresa o	89.646.300-4

institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Subgerente
Rubro, área o sector a la cual se vincula o en la que trabaja	Investigación en empresa de semillas.
23	
Nombre	Rodrigo
Apellido Paterno	Carvacho
Apellido Materno	Baillon
RUT Personal	
Dirección, Comuna y Región	Av El Bosque Norte 0440, piso 8
Fono y Fax	203.51.93
E-mail	rcarvacho@cischile.cl
Nombre de la organización, empresa o institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	Compañía Internacional de Semillas Ltda
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Ing Agrónomo, Jefe programa de Investigación de Remolacha
Rubro, área o sector a la cual se vincula o en la que trabaja	Remolacha, Iansa y empresas productoras de semillas
24	
Nombre	Sybil Amalia
Apellido Paterno	Herrera
Apellido Materno	Foessel
RUT Personal	
Dirección, Comuna y Región	CIMMYT, Apdo Postal 6-641, 06600 México, DF México
Fono y Fax	52-55-5804 2004 ext. 22 46
E-mail	s.herrera@cgiar.org
Nombre de la organización, empresa o	CIMMYT/Swedish University of Agricultural Sciences (SLU)

institución donde trabaja / Nombre del predio o de la sociedad en caso de ser productor	
RUT de la organización, empresa o institución donde trabaja / RUT de la sociedad agrícola o predio en caso de ser agricultor	
Cargo o actividad que desarrolla	Estudiante de doctorado
Rubro, área o sector a la cual se vincula o en la que trabaja	Mejoramiento, fitopatología, investigación

Evaluación de las actividades de difusión

Especificar el grado de éxito de las actividades propuestas, señalando las razones de los problemas presentados y sugerencias para mejorarlos en el futuro. Señalar también las razones por las cuales se hicieron modificaciones al programa propuesto inicialmente, en los casos que corresponda.

Las actividades propuestas inicialmente, que correspondían al programa del curso, se cumplieron en su totalidad.

GOBIERNO DE CHILE
FUNDACIÓN PARA LA INNOVACIÓN AGRARIA

5. EVALUACIÓN DE LA PROPUESTA

Organización durante la actividad (indicar con cruces)[1]

Ítem	Bueno	Regular	Malo
Recepción en país o región de destino según lo programado	X		
Cumplimiento de reserva en hoteles	X		
Cumplimiento del programa y horarios según lo establecido por la entidad organizadora	X		
Facilidad en el acceso al transporte	X		
Estimación de los costos programados para toda la actividad	X		

Evaluación de la actividad de formación

En esta sección se debe evaluar la actividad en relación a los siguientes aspectos:

a) Efectividad de la convocatoria

Buena, ya que se cumplió el objetivo de reunir profesionales de distintas instituciones y empresas, así como alumnos de pre y postgrado.

b) Grado de participación de los asistentes (interés, nivel de consultas, dudas, etc)

Excelente, se mantuvo la asistencia y el interés durante los tres días que duro el curso. Los participantes manifestaron sus dudas e inquietudes permanentemente, las que siempre fueron aclaradas por el profesor.

c) Nivel de conocimientos adquiridos en función de lo esperado (se debe indicar si la actividad contaba con algún mecanismo para medir este punto)

Los conocimientos adquiridos estuvieron a la altura de lo que se esperaba del Dr. José Crossa, quien entregó conceptos complejos y áridos de entender con facilidad y claridad, mostrando la aplicación práctica de estos. El profesor envió un ejercicio a los alumnos con el que evaluará la comprensión y aplicación de los contenidos del curso.

d) Calidad de material recibido durante la actividad de formación

[1] En caso de existir un ítem Malo o Regular, señalar los problemas enfrentados durante el desarrollo de la actividad de formación, la forma como fueron abordados y las sugerencias que puedan aportar a mejorar.

Bueno, ya que reunía conceptos entregados en clases, así como ejercicios que permiten aplicarlos.

e) Nivel de adecuación y facilidad de acceso a infraestructura/equipamiento necesario para el logro de los objetivos de la actividad de formación.

Excelente, se usaron los computadores del Area de Computación de la Facultad de Ciencias Agronómicas, Universidad de Chile, a los que se instaló los contenidos y ejercicios del Profesor. Cada estudiante disponía de un equipo lo que hizo muy fluido el trabajo.

f) Indique las materias que fueron más interesantes, más desarrolladas a lo largo de la actividad de formación y las que generan mayor interés desde el punto de vista de la realidad en la cual se desenvuelve el participante.

Lo más interesante fue el uso de los Biplots y sus variantes como medio de interpretación gráfica de las relaciones e interacciones de los distintos factores que se analizan en los ensayos agronómicos y que representa una herramienta poderosa de análisis.

g) Problemas presentados y sugerencias para mejorarlos en el futuro

En general no se presentaron grandes problemas

Aspectos relacionados con la postulación al programa de formación o promoción

a) Apoyo de la Entidad Patrocinante (cuando corresponda)

__X__ bueno _____ regular _____ malo

Justificar:

b) Información recibida por parte de FIA para realizar la postulación

__X__ amplia y detallada _____ aceptable _____ deficiente

Justificar:

c) Sistema de postulación al Programa de Formación o Promoción (según corresponda)

__X__ adecuado _____ aceptable _____ deficiente

Justificar:

d) Apoyo de FIA en la realización de los trámites de viaje (pasajes, seguros, otros) (sólo

cuando corresponda)

____ bueno __X_ regular ____ malo

Justificar:

e) Recomendaciones (señalar aquellas recomendaciones que puedan aportar a mejorar los aspectos administrativos antes indicados)

Tal vez seria bueno contar con la aprobación con mayor anticipación de manera de realizar las actividades organizativas de manera más holgada.

ANEXO 1
MATERIAL RECOPILADO DURANTE EL CURSO

USER'S GUIDE FOR SPATIAL ANALYSIS OF FIELD VARIETY TRIALS USING ASREML

A. Cadena, J. Burgueño, J. Crossa

Biometrics and Statistics Unit, CIMMYT

Mexico

M. Bänziger

Physiology Unit, CIMMYT

Zimbabwe

A. R. Gilmour

Orange Agricultural Institute

Australia

B. Cullis

NSW Agriculture

Agricultural Research Institute

Australia

February 2000

Preface

The last decade of the millenium has seen major improvements in the options available for the analysis of field trials. Traditionally, the principal method of handling spatial variation in a trial was through the use of incomplete block designs. Experience with many analyses has lead to the realization that spatial variation has multiple sources and block designs often fail to do justice to the spatial variability.

This manual describes an approach to the spatial analysis of field experiments based on the software package ASREML (Gilmour *et al.* 1999). It describes common sources of spatial variation and explains how these can be identified and accounted for in an analysis.

NSW Agriculture makes no warranties with respect to ASREML. Use by CIMMYT of any software does not imply endorsement or recommendation. The user takes full responsibility for all manipulations done with ASREML.

CIMMYT

The International Maize and Wheat Improvement Center (CIMMYT) is an internationally funded, non-profit scientific research and training organization. Headquartered in Mexico, the Center works with agricultural research institutions worldwide to improve the productivity and sustainability of maize and wheat systems for poor farmers in developing countries. It is one of 16 similar centers supported by the Consultative Group on International Agricultural Research (CGIAR). The CGIAR comprises over 50 partner countries, international and regional organizations, and private foundations. It is co-sponsored by the Food and Agriculture Organization (FAO) of the United Nations, the International Bank for Reconstruction and Development (World Bank), the United Nations Development Programme (UNDP), and the United Nations Environment Programme (UNEP).

TABLE OF CONTENTS

1 INTRODUCTION

The main objective of variety trials is to obtain precise estimates of variety means and variety contrasts. Soil fertility, soil water-holding capacity, soil physical characteristics and other environmental factors often vary across an experimental site. Previous history, irrigation, plot trimming, direction of cultivation or harvesting are other man induced sources of variation. Good experimental design can reduce the impact of some of these factors but unless they are appropriately included in the statistical model when they occur, they will result in poor precision in estimates of variety effects and variety contrasts.

We use the term spatial (or nearest neighbour) analysis to refer to an analysis where we investigate the variance structure of each trial and use an appropriate structure for estimation of effects in the trial. This approach does not obviate the need for good experimental design but rather increases it because once a treatment effect is confounded with an environmental effect, the two cannot be disentangled.

ASREML (Gilmour et al., 1999) uses the REML (Residual Maximum Likelihood) estimation method to estimate variance components in the context of mixed linear models. It is a useful tool for analyzing field variety trials as it allows for the fitting of spatial variability within field trials in a variety of ways. It allows for various experimental designs, multiple covariables and performs across site analysis.

This Manual describes how to:

1. install ASREML

2. prepare data files and ASREML command files to perform spatial analyses of replicated and unreplicated variety trials

3. interpret ASREML outputs.

2 BACKGROUND

Spatial variability can be partly controlled by using an appropriate experimental design. Most variety trials use complete or incomplete block designs and are analyzed using the traditional analysis of variance. Block designs attempt an *a priori* reduction of the experimental error considering spatial heterogeneity among blocks. This approach does not consider the presence of spatial variability within blocks, and researchers face the problem of having to find blocks in the field that are homogeneous without knowing their most appropriate shape, dimension and orientation. When field variety trials are laid out in a rectangular array of r rows and c columns with replicates allocated contiguously, then spatial analysis can be performed with the aim of improving precision of estimates of variety effects and variety contrasts.

An appealing idea presented by Papadakis (1937) and developed by Wilkinson et al (1983) is to adjust a plot for spatial variability by using information from the immediate neighbours. One useful measure for examining the heterogeneity patterns of the soil is the spatial autocorrelation of neighbouring plots within rows or within columns. That is form the correlation between residuals at various distances apart. If there is no spatial pattern, all the correlations will be low. If there is pattern in the residuals, neighbouring residuals will be more similar and so have higher correlation. Gleeson and Cullis (1987) proposed to sequentially fit a class of autoregressive-integrated-moving average models (ARIMA) to the plot errors in one direction (rows or columns). This was in the context of randomised complete block experiments. They found that

differencing along the block and then fitting a moving average (MA) correlation structure to the residuals in that direction resulted in big gains in efficiency of the trial. Cullis and Gleeson (1991) extended the previous model to two directions (rows and columns) assuming that, in the field, rows and columns are regularly spaced. However, differencing in this two dimensional analysis was prone to discard treatment information (as shown by Kempton et al, 1994). Grondona et al. (1996) analyzed 35 cereal yield trials using the two-dimensional spatial analysis proposed by Cullis and Gleeson (1991) and found that the autoregressive model in the direction of the rows and columns was the most frequent best model.

Gilmour et al (1997) distinguished between global, natural and extraneous variation. For natural variation arising from unevenness of soil moisture, soil depth or other natural variation, they proposed using a separable autoregressive (AR) correlation structure, without differencing. Thus, they model the natural variation as the direct product of an AR correlation structure for columns and an AR correlation structure for rows, denoted by AR1XAR1. Extraneous variation includes effects introduced by the experimental operations. These operations are usually aligned with rows or columns and are usually modelled with random row and column effects. Global effects include any major (non-stationary) trends across the field. These are fitted as linear trends, cubic smoothing splines, row and column contrasts and covariates.

The variogram is used by Gilmour et al (1997) as a major diagnostic tool for checking for the presence of extraneous variation, along with trellis plots of residuals and plots of other random effects. It is essentially the complement of the spatial autocorrelation matrix but is easier to view and interpret. If there is no pattern to the residual, the variogram is essentially flat. Pattern shows itself in that the variance of differences between residuals which are near to each other will tend to be lower than for those that are from plots far apart. In other words, strong patterns in the variogram indicate that

extraneous variation is present. We will display several variograms pointing out the interpretation of some common patterns. The variogram is used in an informal way. Terms added to the model are then formally tested with F-statistics (fixed terms) or Likelihood Ratio tests (random terms).

The classical approach considers that the response variable Y is modelled by

Y= mu + variety effects + *design effects* + *error*

where the italics denote the random terms. Thus, this model includes a constant term (mu), any covariates, and variety effects as fixed effects. Design or block effects are fitted as random effects to recover between block treatment (variety) information. The random effects are assumed to be independent random variables distributed Normal(0, σ^2) with different variances for blocks and residuals.

The basic spatial model considers that the response variable **Y** is explained by

Y = mu + variety effects + global trend + *design effects* + *error*,

The differences are: the possible inclusion of polynomial trends and other special fixed effects to remove systematic spatial variation, additional terms that are considered for inclusion in the design effects and the provision for the random effects and/or the residuals to be correlated.

An alternative is to assume variety effects are random. This is necessary in unreplicated trials and the multi-site analysis of trials where we wish to have correlations between the performance of varieties in different sites. It is often desirable in two replicate trials. This raises the issue of the difference between treating varieties as fixed and as random. As fixed effects, we obtain the best linear unbiased estimate (BLUE) of the variety effect, that is, the best estimate of the performance of that variety in that trial. Treated as random effects, we obtain the best linear unbiased predictor (BLUP) of the variety effect, that is, the best estimate of the performance of that variety

in future trials. Historically, selection has been based on BLUE estimates and it is then observed that performance after release is usually not as good as obtained in the trials. This is simply because future performance is predicted by the BLUP, not the BLUE.

Gilmour *et al.* (1997) proposed extending the Cullis and Gleeson (1991) approach in a sequential manner. First a two-dimensional separable auto-regressive spatial model of first order (AR1 x AR1) is fitted as the basic spatial model. The AR1 x AR1 model is flexible enough to generally represent many different spatial patterns that arise. They then propose looking at the variogram. If it has the classical AR1 x AR1 appearance and there are no outliers or other obvious problems, this model is accepted. Otherwise, the model is adapted as suggested by the variogram until a reasonable result is obtained. This is an iterative process. It is helpful to discuss the models with the experimenter who knows physical site characteristics and trial management details. They often explain the characteristics of the variogram. These might include variability due to agronomic and experimental practices and procedures such as irrigation flow, sowing and harvesting methods and direction, slope, proximity to roads, trees or rivers, machinery characteristics and site history. These are modelled as spatial covariances, covariates and functions of the spatial coordinates using polynomial functions or cubic splines.

Wald statistics or F ratio statistics can be used to test the significance of fixed effects considered in the model. We prefer to include random terms rather than fixed terms in the model so as to recover treatment information but some effects need to be included as fixed effects. The likelihood ratio test is used to test random effects in the model.

There is often not a single best spatial model but rather, several reasonable spatial models. A reasonable spatial model is one where global, extraneous and natural

sources of variation are included in the model and this will be reflected in a variogram which has little structure other than the basic AR1 x AR1 structure.

A traditional statistic used to test for variety differences is the standard error of difference (SED). This is appropriate for testing differences between fixed (BLUE) variety effects. From spatial analysis, the SED varies for each particular contrast but an average value is often reported. This may be used when all varieties have essentially the same replication. The SED is not an appropriate statistic for choosing a spatial model because it is strongly influenced by the particular spatial model. The best spatial model does not necessarily have the smallest SED. The SED is not appropriate for testing differences in BLUPs, i.e. random effects should be BLUPs.

In general, the approach proposed by Gilmour *et al.* (1997) for applying spatial model to variety trials seems very attractive as it helps researchers to increase the precision of the experiment and have a better understanding on how the data in each particular environment was generated. Their paper should be read as it demonstrates the process of fitting a spatial model to real data. Nevertheless, as pointed out by Brownie *et al.* (1993), the presence of systematic variation within complete of incomplete blocks does not invalidate the use and the analysis of the classical complete block or incomplete block designs (such as lattices) but rather it strengthens the need for the random allocation of varieties to plots within complete or incomplete blocks. In addition improved statistical tools never compensate for precision lost while conducting an experiment using poor design or inappropriate agronomic and experimental practices.

3 INSTALLATION OF ASREML

ASREML is currently available on the internet. A compiled version for personal computers (WINDOWS versions 95, 98 and NT) including manual can be found at the web address: ftp://ftp.res.bbsrc.ac.uk/pub/aar/. It is also available for Sun Solaris and others operating systems. It requires a PC with at least a 486 processor, 16 MB ram, 50Mb hard disk and running WINDOWS® 95 or higher.

This section describes installation of ASREML for personal computers and assumes that the program file ASRWIN.EXE has been downloaded to a floppy disc, or that the user uses a floppy disc provided with this manual. Installation details are contained in the file INSTALL.TXT. Briefly they are as follows:

1. Go to **MSDOS Prompt**
2. Type **MKDIR C:\ASREML** and press ENTER (or make a subdirectory C:\ASREML using WINDOWS Explorer)
3. Type **CD C:\ASREML** and press ENTER
4. Place the floppy disk in Drive A and type **A:ASRWIN** and press ENTER

With these instructions, the program files are UNPACKED. It is highly recommended to run a trial to verify that the program is functioning properly. For this, proceed as follows:

6. Type **ASREML SHF** and press ENTER

SHF is an ASREML instructions file generated when the program is UNPACKED. If ASREML has been successfully installed, a variogram that is part of the ASREML output will appear. You can scroll through the outputs and leave ASREML by consecutively pressing the ENTER key.

ASREML memory requirements vary with the size of the job. Typically, analyses of a single trial will require at least 10MB hard disc space. If there is not enough memory or hard disc space, an error message describing the problem will pop up (**INSUFFICIENT**

MEMORY) or a reference will be made to the **error number 169** or to the file **LF90.EER**, which contains the text associated to the FORTRAN errors.

A complete manual for ASREML by Gilmour *et al.* (1999) in Postscript format can be obtained by downloading the archive file ASRPS.EXE (available at ftp://ftp.res.bbsrc.ac.uk/pub/aar/) and running it to produce ASREML.PS. If you do not have a Postscript printer, you can read and print the manual using the program ghostscript (available at http://www.cs.wisc.edu/~ghost/index.html).

4 THE LINEAR MIXED MODEL

Formally, ASREML estimates variance components in a general linear mixed model using the residual maximum likelihood (REML) approach. The equation for the general mixed linear model is:

$$\mathbf{Y} = \mathbf{X}\tau + \mathbf{Z}\mathbf{u} + \eta \qquad (1)$$

where

- $\mathbf{Y}$ is the response vector
- $\mathbf{X}$ is the design matrix for fixed effects
- τ is the vector for fixed effects
- $\mathbf{Z}$ is the design matrix for random effects
- $\mathbf{u}$ is the vector for random effects
- η is the vector of residuals

Matrix $\mathbf{X}$ is the design matrix for fixed effects such as overall mean and varieties (τ). Matrix $\mathbf{Z}$ is the design matrix for random effects ($\mathbf{u}$) such as complete or incomplete blocks, row and column effects, splines in the direction of the rows and/or columns.

Here one can also include the extraneous effects due to agronomic practices and other experimental procedures. The residual (η) is composed of:

I. the local trend (ξ) which is modeled by the two dimensional auto-regressive procedures in the direction of the row and columns, and

II. the residual **e** after adjusting for all the other terms in the model.

Note that $\eta = \xi + \mathbf{e}$. The random terms (**u**, ξ, **e**) are pairwise independent.

ASREML fits this model with **u** and ξ or **e** having zero mean and variance-covariance matrices given by Var(**u**) = **G** and Var(η) = **R**. The default cases are when $\mathbf{R}=\sigma^2\mathbf{I}$ and $\mathbf{G}_f=\sigma_f^2\mathbf{I}$.

5 DATA ANALYSIS USING ASREML

This section provides a general outline for using ASREML. The examples in Section 6 will help the user to comprehend this outline.

5.1 DATA FILE

Data of field trials is commonly organized using a spreadsheet. Columns of the spreadsheet represent the distinct attributes of each plot (factors, variables, covariables). Data for each experimental unit are listed in rows. For ASREML to read a data file properly, it should be saved as a comma separated file (.CSV). If not prepared in a spreadsheet, the file should be prepared as an ASCII text file with columns separated by spaces.

Characteristics of the data file are:

- Identifiers (headings, titles) of columns may be included.

- Identifiers (headings, titles) and attributes (data, factor levels) must be alphanumeric.
- Columns must be separated by at least one blank space, TAB or comma (in .CSV file).
- Missing data must be represented by a dot/period (.), an asterisk (*) or NA. In a .CSV file, empty fields between commas are considered as missing values; a row that begins or ends in a comma is considered as having missing data at the beginning or at the end, respectively, of that row.
- If a row has insufficient data fields, the input record is completed by taking values from the following row. This may result in wrong allocation of data.
- A number sign (#) and dollar sign ($) have special meanings. Neither may appear in the data file.

In .csv format, a typical data file might begin like

```
Env,Expt,Rep,Row,block,check,plot,entry,yield,Column
4,14,1,1,1,0,8601,8601,3.47,1
4,14,1,1,2,0,8602,8602,3.58,2
4,14,1,1,3,0,8603,8603,4.17,3
4,14,1,1,4,1,8604,8604,3.87,4
```

In space delimited ascii text form, it might look like

```
Env Expt Rep Row block check plot entry yield Column
  4   14   1   1     1     0  8601  8601  3.47      1
  4   14   1   1     2     0  8602  8602  3.58      2
  4   14   1   1     3     0  8603  8603  4.17      3
  4   14   1   1     4     1  8604  8604  3.87      4
```

5.2 COMMAND FILE

ASREML uses a command file that defines data file, attributes, model and variance structure. The following rules apply:

- The file must have a valid DOS name and the extension <name>.AS (for example: TRIAL1.AS). Avoid names with embedded spaces.
- The command file consists of five sections. These are:
  I. Title
  II. Definition of data columns
  III. Name of the data file
  IV. Linear model
  V. Variance structure (when necessary)
- All characters after a number sign (#) are ignored
- Blank rows are ignored.
- A blank space is the most commonly used separator, although a TAB may also be used.
- Key words are sensitive to capital and lower-case letters.
- All rows starting with a ! followed by a blank space are copied as comments to the output file.

Example of command file to analyse a single trial

```
Spatial analysis of a field trial          #  TITLE
 env expt                            # DEFINITION OF COLUMNS
 rep 2
 row 8 block 8 check
 plot
 entry 64 !I
 yield
 column 16
#                          Some data as Comments
#  4 14  1  1  1  0    8601    8601    3.47    1
#  4 14  1  1  2  0    8602    8602    3.58    2
#  4 14  1  1  3  0    8603    8603    4.17    3
#  4 14  1  1  4  1    8604    8604    3.87    4
#  4 14  1  1  5  0    8605    8605    2.92    5
```

```
trial.dat !SKIP 1          # Name of data file; skip header
line
yield ~ mu entry           # Linear model

1 2                        # Variance structure  AR x AR
row row AR .1
col col AR .1
```

In this guide, we present the information required for coding ASREML for spatial analyses. Much more information and many more options are in the reference manual.

I Title

The first line in the command file is used for the title. It should clearly identify the trial for future reference to this analysis.

II Definition of columns in the data file

Each column of the data file must be defined. The definition consists of a <name>. Factors are defined by also specifying the number of <levels> and usually a <qualifier>. Column definitions must be **indented.** While several definitions may appear on a line, it is less confusing to have one definition per line.

- **<name>** A column name may have up to 20 characters preceded and followed by a blank space.
- **<levels>** If the column contains a variable, <levels> is set to 1 or omitted; if the column contains a factor, <levels> is set to 2 or the actual number of factor levels.
- **<qualifiers>** The major qualifiers are **!A** and **!I**.

  !A means that the data in this column is alphanumeric. ASREML will recode them in the order of appearance.

!I means that the data in this column is numeric but not 1 -- <levels>. ASREML will recode them in the order of appearance.

No qualifier: If <levels> is greater than 1 but no qualifier used, the data in this column is considered as a factor coded 1 to <levels>. Otherwise the values are considered as a variable.

ASREML has some capacity to transform data which is described in the reference manual. We recommend new users transform data in a spreadsheet before using ASREML.

III File name for the data file and general qualifiers

A data file must always be specified after defining the data columns. Its name must begin in the first position of the line. The data file name must include the path if it is not in the same directory as the command file.

There are many qualifiers that can be placed on this line after specifying the name of the data file. The most commonly used are

!skip *n* indicates to skip over the first *n* lines (those containing column headings) in the data file

!maxit *m* establishes the maximum number of iterations in *m*. The default is 10 iterations.

IV Linear model

The linear model is a list of terms, each separated by a space in the form

<variable Y> ~ <model>

<variable Y> is the name of the data field which will be analyzed.

<model> lists the terms of the model. <variable Y> is separated from <model> using the symbol (~).

Some common model terms are:

mu	represents a constant term or the intercept
<name>	is the name of an explanatory variable or factor
<name>.<name>	is the interaction of two terms
!r	indicates that the following term is *random*
!f	indicates that the following term is *fixed*
mv	if there are missing values in the response variable <variable Y>, **mv** needs to be placed among fixed terms i.e. before **!r** or after **!f**.

Examples of models are:

Y ~ mu entry

Yield ~ mu entry !r block

Height ~ mu con(entry) lin(col) !r spl(col)

con(), lin() and spl() are model functions described below.

Some rules when writing the model

- Terms in the model are sensitive to capital and lower-case letters.
- Wherever possible use the full name of factors and covariates when listing the model terms as this avoids confusion. However, it is legal to truncate names provided ambiguity is avoided.

- ASREML provides several model functions to achieve particular forms in the design matrix. The major ones are:

  con(entry) in place of *entry* constrains *entry* effects to sum to zero,

  lin(col) in place of *col* treats the factor *col* as a covariate, and

  pol(col,t) forms a t-orthoganal polynomials from co; the mean is excluded if it is negative, *pol(col,-t)*.

  spl(col), used as a random term in conjunction with *lin(col)* as a fixed term, fits the cubic smoothing spline to col.

  con() lin(), and pol() should only be used in the fixed part of the model. spl() should only be used in the random part

- Interactions between factors can be simplified. For example

  *a*b* expands to *a b a.b*

  a.(b c d) e expands to *a.b a.c a.d e*

V Variance model

A random term has a variance structure associated with it. The default variance structure is independent (uncorrelated) effects with equal variance. For the models covered in this manual, it is sufficient to assume the effects in all random factors other than the residual are uncorrelated and that the residuals may be correlated with a separable autoregressive structure. ASREML can handle much more complicated situations.

Random effects: Uncorrelated random effects are included in the linear model by listing them after **!r**. These commonly include factors like *row*, *column*, *block*, *rep*, *spl(row)*, *spl(col)* and some interactions.

Residuals: The residuals may be modelled as distributed independently or with a separable autoregressive structure [AR(rows) x AR(columns)] and occasionally as a

combination of these. For the first case, nothing extra is required. For the second case, three extra lines are required. Typically they will be

```
1   2
column column AR1  .1
row row AR1  .1
```

These lines mean:

- that we are analysing a single experiment laid out in the field in two dimensions;
- that the data file contains two factors called *row* and *column* which index the field positions and
- that ASREML is to fit an autoregressive structure to each of these dimensions using 0.1 as the initial correlation.

With this specification, ASREML checks that the spatial arrangement is correct.

Example of randomised block analysis

A randomised block analysis could be specified in ASREML with:

Command file	***Explanation***
Example: randomised block analysis	*Title*
rep 3	*Column definition*
entry 107	*Column definition*
yield	*Column definition*
row 22	*Column definition*
col 15	*Column definition*
mydata.dat !SKIP 1	*File name for data file*
y ~ mu con(entry) !r rep	*Linear model*

The data file has 1 header line and 330 data lines. Each data line contains a rep number (1—3), a genotype (entry) number (1—107), a yield, a row number (1—15) and a column number (1—22). The spatial information is ignored in this analysis.

The **fixed model** includes the general mean [mu] and entry effects with the restriction that the sum of all entry effects equals zero [con(entry)]. Replicate effects are fitted as random effects with variance $var(\mathbf{u})=\sigma^2\gamma_r\mathbf{I}_3$ and residuals are distributed with variance $var(\mathbf{e})=\sigma^2\mathbf{I}_{330}$.

Example: Initial spatial analysis

The initial spatial analysis as advocated by Gilmour et al (1997) could be coded as:

Command file	***Explanation***
`Example of spatial analysis`	*Title*
`rep 3`	*Column definition*
`entry 107`	*Column definition*
`yield`	*Column definition*
`row 22`	*Column definition*
`col 15`	*Column definition*
`mydata.dat !SKIP 1`	*File name for data file*
`y ~ mu c(entry)`	*Linear model*
`1 2`	*Variance structure*
`col col AR1 0.1`	*Variance structure*
`row row AR1 0.1`	*Variance structure*

The **fixed model** is as before, replicates are not fitted but the residuals are distributed with variance $var(\eta)=\sigma^2\Sigma_A(\phi_{Col}) \otimes \Sigma_A(\phi_{Row})$ where $\Sigma_A(\phi)$ is a matrix of auto-regressive correlations, with the parameter ϕ corresponding to columns and rows, respectively. Correlation parameters between 0 and 1 are valid.

5.3 RUNNING THE PROGRAM

The command to run ASREML is

C:\ASREML\ASREML.EXE -<option> <command file>

where common <option>s include c, m, n and s2 and <command file> is the name of the command file. This command may be typed directly at the command

prompt in an MSDOS box, may be associated with the .as filename extension within Explorer or entered after clicking **Start Run**

Concerning the options, use:

-c to resume a run continuing iterations from the current point ;

-m to invoke ASREML's Menu for interactive running and viewing output,

-n to suppress the graphics and

-s2 to increase memory size.

-g11 to save the graphics to WMF files.

Multiple options should be concatenated. E.g. -cms2 to combine c, m and s2.

5.4 RESULTS

The results are written in various output files. A detailed summary of the analysis appears in the file <name>.ASR. The file <name>.SLN includes adjusted effects. In both cases, <name> is the same as the command file.

6 EXAMPLES FOR SPATIAL ANALYSIS OF A VARIETY TRIAL

All data and command files used in this manual can be found in the following internet address www.cimmyt.cgiar.org/biometrics

6.1 SPATIAL ANALYSIS OF A VARIETY TRIAL WITH REPLICATES

This section presents a sequence of programs in ASREML used for spatial analysis of replicated field trials. Data from two trials will be shown (TRIAL1.DAT and TRIAL2.DAT). Note that the appropriate use of ASREML for spatial analysis of field trials is not automatic and can not be done by simply pressing a couple of keys. As proposed by Gilmour et al (1997), the analysis proceeds through a series of steps to produce an appropriate spatial model.

6.1.1 TRIAL 1

I Preparing the data file

Our first example is a variety trial designed as an alpha-lattice design (Patterson and Williams, 1976) with 16 varieties planted in three contiguous replicates laid out in 6 rows and 8 columns. Each replicate had 2 rows. Each row had 2 blocks of 4 plots. Data for this example are in file TRIAL1.DAT. The first rows of the data file are shown below.

First rows of the data file TRIAL1.DAT:

rep	blk	row	col	plot	Variety	yld
1	1	1	1	1	16	2556
1	1	1	2	2	1	1361
1	1	1	3	3	6	1567
1	1	1	4	4	10	1797
1	2	1	5	5	11	2753
1	2	1	6	6	4	2089
1	2	1	7	7	15	2531
1	2	1	8	8	5	3144
1	3	2	1	9	12	2189
1	3	2	2	10	3	1864

```
1    3    2    3    11      2   1400
1    3    2    4    12      7   1006
...  ...  ...  ...  ...     ...  ...
```

Each column or data field is separated by blank spaces and represents an attribute (factor, variable). The first line is a header line. There is then a line for each experimental unit, or plot. Thus the second line indicates that this experimental unit is from replicate (rep) 1, block (blk) 1, row 1, column (col) 1, plot 1, variety 16 and has a grain yield (yld) of 2556 kg ha^{-1}.

The usual first model to fit following the approach of Gilmour et al (1997) is assume separable AR1 x AR1 correlated errors. However, to introduce the coding of ASREML, first present the traditional RCB (randomized complete block) analysis and the incomplete block analysis.

II Analysis as a randomized complete block design (RCBD)

The command file, RCBD.AS, contains the instructions for analyzing the data in TRIAL1.DAT as a randomized complete block design.

Command file RCBD.AS for analyzing TRIAL1.DAT as a randomized complete block design:

```
Randomized complete block analysis with ASREML
 rep  3
 blk  12
 row  6
 col  8
 plot 48
 var 16
 yld
trial1.dat  !skip 1
yld ~ mu var !r rep
```

This file has four sections:

Title: Randomized complete block analysis with ASREML

Definition of columns: The column definitions must all start with a blank space.

- rep is a factor with 3 levels.
- blk is a factor with 12 levels coding blocks within replicates
- row is a factor with 6 levels.
- col is a factor with 8 levels.
- plot is a factor with 48 levels.
- var is a factor with 16 levels.
- yld is the dependent variable

Name of the data file: trial1.dat !skip 1

!skip 1 causes ASREML to ignore the first line of the data file since it is a header line.

Linear model: The mean (mu) and variety are fixed effects. rep is declared to be a random effect by placing it after !r.

Results: ASREML generates several output files. The primary output file, with file extension .asr, for the complete block analysis follows. A detailed description of the contents of this file is given with the ASREML output from the incomplete block analysis. The Loglikelihood from this RCB analysis is REML logl=-223.482.

Output file RCBD.ASR for TRIAL1.DAT analyzed as a randomized complete block design:

```
ASREML [2 Sep 1999]  Randomized complete block analysis with ASREML
28 Sep 1999 14:46:10.450    8.00 Mbyte     rcbd
Reading trial1.dat  FREE FORMAT skipping  1  lines
Univariate analysis of yld
Using      48 records [of      48 read from      48 lines of trial1.dat  ]
Model term      Size Type     COL   Minimum    Mean      Maximum   #zero #miss
  1 rep            3 Factor    1         1     2.0000          3         0
0
  2 blk           12 Factor    2         1     6.5000         12         0
0
  3 row            6 Factor    3         1     3.5000          6         0
0
  4 col            8 Factor    4         1     4.5000          8         0
0
  5 plot          48 Factor    5         1    24.5000         48         0
0
```

```
  6 var            16 Factor    6      1    8.5000          16          0
0
  7 yld             1 Variate   7  228.0     1331.        3144.         0
0
  8 mu              1 Constant Term
Forming  20  equations:  17  dense
 Initial updates will be shrunk by factor    0.548
 NOTICE:      1 (more) singularities,
 LogL=-226.562     S2= 0.28241E+06     32 df    0.1000       1.000
 LogL=-224.962     S2= 0.24775E+06     32 df    0.2040       1.000
 LogL=-223.895     S2= 0.22317E+06     32 df    0.4255       1.000
 LogL=-223.547     S2= 0.21198E+06     32 df    0.7227       1.000
 LogL=-223.486     S2= 0.20767E+06     32 df    0.9627       1.000
 LogL=-223.482     S2= 0.20659E+06     32 df     1.048       1.000
 Final parameter values                          1.0556      1.0000

Source              Model  terms     Gamma      Component   Comp/SE   % C
 rep                    3      3   1.05562         218080.      0.95   0 P
 Variance              48     32   1.00000         206588.      3.87   0 P

 Analysis of Variance        DF     F-incr     F-adj  StndErrDiff
  7 mu                        1      23.15      6.99
  5 var                      15       2.29      2.29        371.1

Solution      Standard Error     T-value     T-prev
  5 var
                 2    -295.000      371.114      -0.79
                 3     494.667      371.114       1.33        2.13
                 4     825.333      371.114       2.22        0.89
               ...         ...          ...        ...         ...
                14     216.000      371.114       0.58        0.51
                15     886.333      371.114       2.39        1.81
                16     336.333      371.114       0.91       -1.48
7 mu
                17     993.333      375.580       2.64
  1 rep                    3 effects fitted

Finished: 28 Sep 1999 14:46:13.750   LogL Converged
```

III Analysis as a randomized incomplete block design (RIBD)

The command file BLOCK.AS contains the instructions for analyzing the data in TRIAL1.DAT as an incomplete block design.

Command file BLOCK.AS for analyzing TRIAL1.DAT as an incomplete block design:

```
Incomplete block analysis
 rep 3
 blk 12
 row 6
 col 8
 plot 48
 var 16
 yld
trial1.dat !skip 1
```

```
yld ~ mu var !r rep blk
```

The only difference to the previous program (RCBD.AS) is the addition of the random factor blk. If the blocks were coded 1...4 within reps rather than 1...12 across reps, the block factor would need to be fitted as rep.blk.

Results: The primary ASREML output has 6 sections which are discussed in detail. It is important to understand the output to avoid accepting invalid analyses.

Section 1: The first line displays the compilation date of the program and the title line for the job. The second line displays the date and time of the run, the size of the data space being used and the name of the job being run. The third line gives the name of the data file and number of header lines being skipped. The fourth line names the dependent variable. The next line indicates how many data records have been read and how many are being used in the analysis.

```
ASREML [ 2 Sep 1999]  Incomplete block analysis
 05 Oct 1999 14:52:35.370   8.00 Mbyte block.as
 QUALIFIERS: !skip 1
 Reading triall.dat  FREE FORMAT skipping  1 lines
 Univariate analysis of y
Using      48 records [of      48 read from      48 lines of triall.dat     ]
```

Section 2: This contains a summary of the data. Things to check here are that the labels for the terms match the data values, that the ranges, number of zeros and missing values are correct. An idiosyncrasy of ASREML is that the Minimum is determined ignoring zeros since these are reported in the #zero column.

Model term	Size	Type	COL	Minimum	Mean	Maximum	#zero	#miss
1 rep	3	Factor	1	1	2.0000	3	0	0
2 blk	12	Factor	2	1	6.5000	12	0	0
3 row	6	Factor	3	1	3.5000	6	0	0
4 col	8	Factor	4	1	4.5000	8	0	0
5 plot	48	Factor	5	1	24.5000	48	0	0
6 var	16	Factor	6	1	8.5000	16	0	0
7 yld	1	Variate	4	228.0	1331.	3144.	0	0
8 mu	1	Constant	Term					

Section 3 reports information produced while analyzing the model. The 32 equations refer to the order of the mixed model equations. That is 1 + 16 + 3 + 12. These are divided into a 'dense' set, the first 17 and a 'sparse' set, the 15 random effects. The update factor limits the step size of the parameter updates in the first few iterations. This is a strategy to facilitate convergence.

Singularities are linearly dependent equations. One singularity occurs because ASREML cannot estimate, without a constraint, 16 variety effects and a mean. The constraint used is to fix the first variety effect to zero such the overall mean effect, **mu,** is actually the mean of variety 1 and that the variety effects are actually deviations from variety 1.

Next is a report of the iteration process. The logl value is the Loglikelihood which increases to a maximum of REML LogL=–221.417. S2=0.13322E+06 is the converged estimate of the residual variance. There is then a statement of the number of residual degrees of freedom (32). This is a maximum value to use in testing any F-statistics shown below. This is followed by the values of the variance parameters at each iteration.

```
Forming  32  equations:  17  dense
 Initial updates will be shrunk by factor     0.548
 NOTICE:    1 (more) singularities,
 LogL=-221.417     S2= 0.13322E+06     32 df     1.508     0.6523     1.000
 LogL=-221.417     S2= 0.13322E+06     32 df     1.508     0.6523     1.000
 Final parameter values                          1.5077    0.65229    1.0000
```

Section 4 summarizes the analysis. The first table presents the variance components. The variance components are derived from the Gamma values (used in the iteration) by multiplying them by the Variance (S2).

The Comp/SE is similar in concept to a t-statistic and provides a measure of the size of the component. It is usual to test variance components with a LikeLihood Ratio test (LRT), i.e. test the decrease in the REML logl value produced by dropping the term

from the model. The LRT is usually not significant if Comp/SE is less than 0.5 and is usually significant if Comp/SE is greater than 1.5.

The % column is the percentage change in the parameter in the final iteration. At convergence it will be 0. The final column indicates any parameter constraints. A "P" in this column indicates the parameter is in parametric space, a "B" indicates the parameter has been fixed at a boundary. An "S" in this column indicates the variance model is over-parameterized and there is no information to update the parameter.

The analysis of variance table provides tests of fixed effects. Two F ratios are presented. The first is like the SAS Type I test. It tests the addition of this term in the model after adjusting for all effects not in the table or higher in the table. The second F ratio is like the SAS Type III test. It tests the term after adjusting for all other terms in the model. The Type III test is meaningless for some terms when there are singularities and interactions in the model. ASREML also reports an average Standard Error of Difference (SED) for main effects when the model is simple.

Source	Model	terms	Gamma	Component	Comp/SE	% C
rep	3	3	1.50773	200856.	0.87	0 P
blk	12	12	0.652285	86895.9	1.33	0 P
Variance	48	32	1.00000	**133218.**	3.26	0 P

Analysis of Variance	DF	F-incr	F-adj	StndErrDiff
7 mu	1	23.01	6.48	
5 var	15	2.75	2.75	**326.4**

Section 5 shows the overall mean (**mu**) and solutions for the fixed factors. For this simple model, variety means are calculated by adding **mu** to the variety effects. For example, for variety 1 the adjusted mean is 907.978, for variety 2 the adjusted mean is the sum of mu (907.978) and the effect for variety 2 (-284.283), giving an adjusted mean of 623.695. All effects in the model are listed in the .sln file. For more complicated situations you will need to refer to chapter 6 of the ASREML reference manual to see how you might form linear functions of these solutions. T-value and T-prev, test hypothesis of effect equal zero and differences between consecutive effects respectively.

		Solution	Standard Error	T-value	T-prev
5 var					
	2	-284.283	333.736	-0.85	
	3	493.858	335.587	1.47	2.42
	4	699.294	320.937	2.18	0.64
	5	863.399	323.003	2.67	0.51
	6	30.9743	309.614	0.10	-2.49
	7	-115.815	323.003	-0.36	-0.44
	8	157.764	333.736	0.47	0.82
	9	407.242	333.600	1.22	0.78
	10	777.177	320.937	2.42	1.15
	11	1038.46	333.600	3.11	0.81
	12	648.027	335.587	1.93	-1.21
	13	248.187	320.937	0.77	-1.24
	14	405.458	322.862	1.26	0.49
	15	942.484	333.600	2.83	1.60
	16	453.785	320.937	1.41	-1.52
7 mu					
	17	907.978	356.640	2.55	
1 rep		3 effects fitted			
8 blk.rep		12 effects fitted			

Section 6 gives the time and date when the analysis was concluded and indicates if there was convergence in the iterative process, both, log-likelihood and parameters estimates. Three termination messages are common:

LogL converged indicates that the iteration process has converged satisfactorily. It occurs when LogL difference in two consecutive iterations is less than 0.002 and the variance parameters are also not changing.

WARNING: LogL Converged; Parameters Not Converged indicates the LogL values are very close but the variance parameters are still changing. The percentage change of each parameter in the last iteration is reported in the % field of the variance component report. You may choose to accept the result or force a few more iterations by rerunning the job using the -c command line option [e.g. ASREML -c BLOCK] which will do an extra 2 or 3 iterations. This has been the case here and other examples of this manual.

Warning: LogL not converged means you should review the job. If it was converging and just needs a few more iterations, you can rerun the job with the **!CONTINUE** qualifier and/or using the **!MAXIT** qualifier to request more iterations. Both are placed

on the data-file-name line. **!CONTINUE** causes the analysis to resume with the results from the most recent iteration. Putting say **!MAXIT 20** increases the number of iterations from the default 10 to 20. The -c command line option is equivalent to specifying **!CONTINUE**. If the LogL is erratic look for some explanation, simplify the model and rerun.

```
Finished: 05 Oct 1999 14:52:37.950   LogL Converged
```

IV Spatial analysis using the AR1 x AR1

The spatial modeling approach of Gilmour et al (1997) begins by fitting the AR1 x AR1 error model and looking at the residuals and a variogram of them in a graph. The AR1 x AR1 model fits the natural local variation well. The plot of residuals often reveals anomalous points and global trends in the data. Use the following ASREML program:

Command file AR1AR1.AS for analyzing TRIAL1.DAT using the AR1 x AR1 model:

```
Alpha lattice example
 rep  3
 blk  12
 row  6
 col  8
 plot  48
 var  16
 yld
trial1.dat !skip 1
yld ~ mu var
1 2
row row AR1 0.1
col col AR1 0.1
```

This model does not include any random terms other than the residual. There are three lines on the end which define the structure of **R**, the covariance structure for the residuals. **1 2** indicates that the analysis is conducted for one site laid out in two dimensions. The next two lines define the two dimensions, associating them with the factors row and col and indicating that an AR model is to be fitted to both dimensions. The double occurrences of row and col have specific meanings. The first is used to declare the size of the dimension, i.e. how many rows (columns) are present and depends on the row (col) factor being declared as a factor with the correct number of rows (columns). The actual number of rows (columns) could be inserted into the first position rather than referring to the factor. The second is used to control the order of the records. ASREML needs to know the field order. If the second field is zero, the data is assumed presented with, in this case, columns nested within rows. Naming the respective row and column factors causes ASREML to sort the records so that there is no chance of confusion on this count. It is therefore recommended that row/column

coding be included in the data and that they be used in this way to ensure the correct field order is assumed.

The AR1 0.1 coding indicates that a first order autoregressive correlation structure with initial correlation value of 0.1 is required. The main alternative used in spatial modeling is the ID coding. Note that no parameter value is required for the identity in this context.

Partial output file AR1AR1.ASR for TRIAL1.DAT analyzed using the AR1 x AR1 model

```
     6  AR=AutoR     0.10
     8  AR=AutoR     0.10
Forming  17  equations:  17  dense
 Initial updates will be shrunk by factor     0.548
 NOTICE:      1 (more) singularities,
 LogL=-229.632      S2= 0.36929E+06     32 df     1.000      0.1000      0.1000
 LogL=-223.070      S2= 0.31100E+06     32 df     1.000      0.2784      0.5000
 LogL=-221.639      S2= 0.36032E+06     32 df     1.000      0.2616      0.6811
 LogL=-221.490      S2= 0.39375E+06     32 df     1.000      0.2339      0.7321
 LogL=-221.475      S2= 0.40786E+06     32 df     1.000      0.2271      0.7476
 LogL=-221.473      S2= 0.41303E+06     32 df     1.000      0.2246      0.7527
 Final parameter values                           1.0000     0.22376     0.75451

 Source                Model  terms      Gamma      Component     Comp/SE   % C
 Variance                 48     32    1.00000         413035.        2.78   0 P
 Residual           AR=AutoR      6   0.223765       0.223765        0.79   0 U
 Residual           AR=AutoR      8   0.754510       0.754510        7.67   0 U

 Analysis of Variance       DF      F-incr      F-adj  StndErrDiff
   7 mu                      1       46.79      13.40
   5 var                    15        2.63       2.63        296.1
```

We note first that the REML logl is –221.473, 2.01 higher than the –223.482 obtained from the complete block model, and 0.05 lower than (almost the same as) the -221.417 obtained with the incomplete block analysis with the same number of parameters.

When conducting a spatial analysis with a two-dimensional structure to errors, ASREML generates two graphs of interest:

- trellis plot of the residuals with marginal means. (**Figure 1, left**) where the marginal diamonds indicate the mean, minimum and maximum of the residuals in the row (right side) or column (top) and
- the variogram of the residuals (**Figure 1, right**) where rows are plotted on the left

side and columns are on the right side.

The variogram is a fundamental tool that guides us in improving the basic AR1 x AR1 spatial model. However, there are no formal tests or procedures associated with these displays. We formally test terms suggested by these figures using likelihood ratio tests (for random effects) or approximate F tests (fixed effects) and limit our selves to terms for which there is a plausible biological basis. Common terms are strong (non-stationary) trends across the experiment, edge effects and row/column effects probably induced by agronomic processes associated with conducting the trial (serpentine sowing or harvesting, unequal plot sizes, machinery effects).

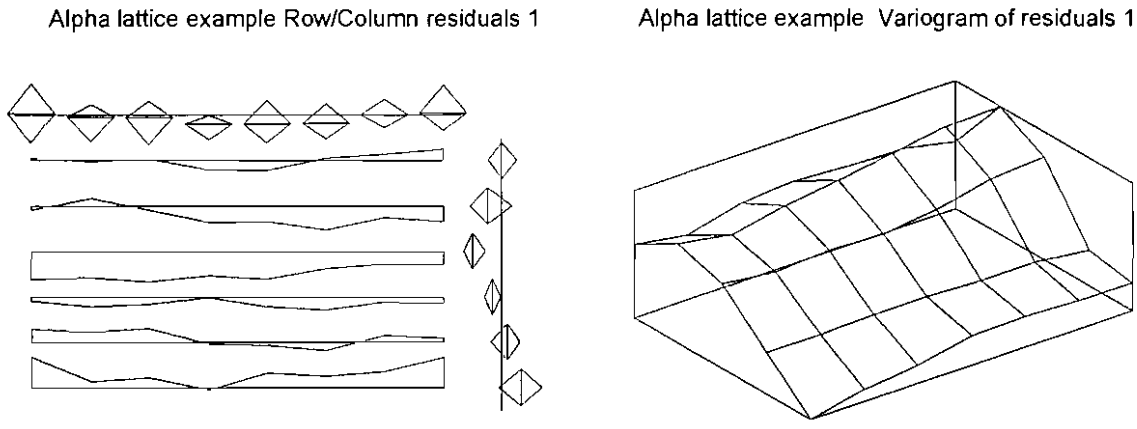

Figure 1 *Trellis plot (left) and Variogram (right) of residuals from the AR1 x AR1 model produced by ASREML. Row distance is on the left axis, column distance is on the right axis.*

We first note from Fig 1 that there are no obvious points which might be outliers. Such points would stand out with huge diamonds in the margin pointing to them.

Second, we note the variogram is quite smooth with strong effects which do not have the typical AR1 x AR1 appearance. In particular residuals in the same row have much less variation than residuals in the same column. *I.e.* there are strong row effects. This

suggests there are non-stationary trends present. The trellis plot margins suggest that these might be strong curvature associated with the row means of the residuals and weaker curvature associated with the column means of residuals. The typical pattern if there was just autoregressive spatial variation is to have a generally flat pattern except at low row and column lags.

There are two approaches for fitting such curvature. The traditional approach is to fit polynomials. Another approach is to use cubic smoothing splines. We prefer the latter because it is a non-parametric curve. Both models fit the same linear component. They only differ in the way the curvature component is fitted. An advantage of the spline model is that it allows for recovery of treatment information from the curvature component because it is fitted as a random effect. The quadratic model does not provide such recovery because is fitted as a fixed effect.

V. Spatial analysis using the AR1 x AR1 plus extensions

Command file for analyzing TRIAL1.DAT using the AR1 x AR1+pol(row,2) model:

```
Alpha lattice example
 rep 3
 blk 12
 row 6
 col 8
 plot 48
 var 16
 yld
trial1.dat   !skip 1
yld ~ mu var pol(row,-2)
1 2
row row AR1 0.1
col col AR1 0.1
```

This analysis produces the trellis plot of the residuals in Fig. 2a and the variogram depicted in Fig. 2b. The F ratio for pol(row,-2) is 13.38 ($P < 0.01$) and it can be seen that the major non-stationarity along rows in Fig 1 has been removed in Fig 2a. The REML logl for this model *is* –203.69 (Table 1). However, the trend along the columns appears stronger now, because we have removed row effects, indicating we may also

need to fit pol(col,-2). The command file (not shown) for this model just includes *pol(col,-2)* right after *pol(row,-2)*.

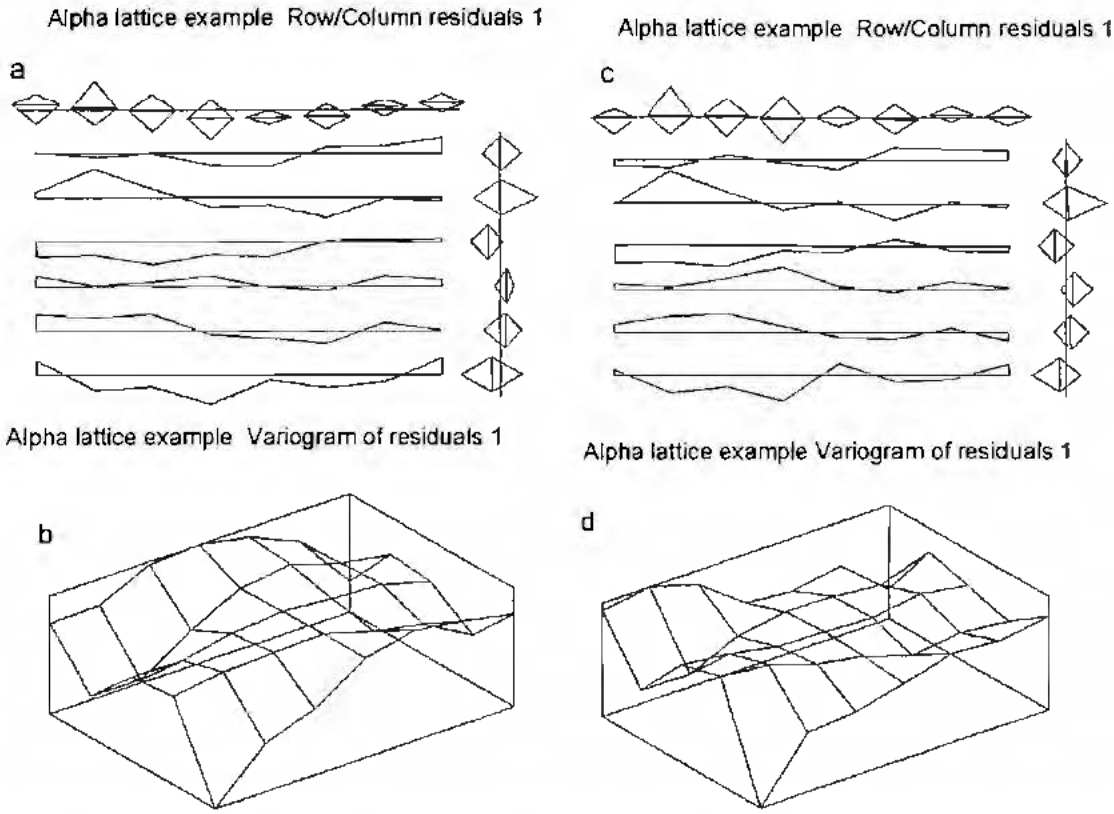

Figure 2 *Trellis plots (above) and variogram plots (below) for spatial models for:* (**a**) *and* (**b**) `AR1 x AR1+ pol(row,-2)` *and* (**c**) *and* (**d**) `AR1 x AR1+ pol(row,-2) + pol(col,-2)` *as fixed terms in the model.*

The F ratio for pol(col,-2) is 5.01 (P<0.05) So, after adding a quadratic trend for both rows and columns the log-likelihood is -189.61. Note that the addition of fixed effects is tested by F-ratios and not be the comparison of log-likelihood values. Trellis plot and variogram of that model are shown in Figs. 2c and 2d.

A cubic smoothing spline is fitted in ASREML by including the terms lin(row) lr spl(row). The lin(row) term is an alternative to pol(row,-1) which does not centre or rescale the

variable fitted. The spl(row) term fits the curvature. The commands file for the final model is

```
Alpha lattice example
 rep 3
 blk 12
 row 6
 col 8
 plot 48
 var 16
 yld
trial1.dat !skip 1
yld ~ mu var lin(row) lin(col) !r spl(row) spl(col)
1 2
row row AR1 0.1
col col AR1 0.1
```

Note that the terms spl(col) and spl(row) are written after **!r** and thus are considered as random effects.

The plots from this model are shown in Figure 3 and are very similar to those for the two dimensional polynomial model.

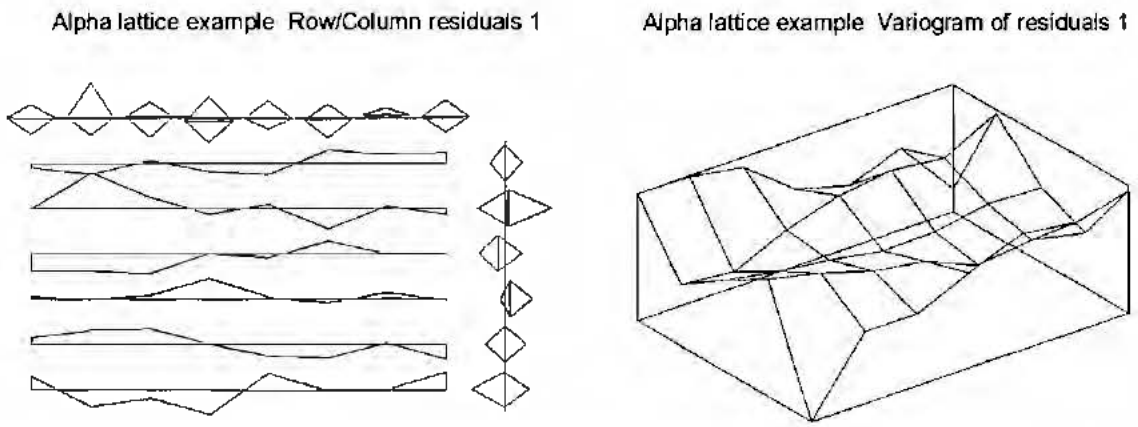

Figure 3 *Trellis plot (left) and variogram plot (right) for spatial models with* `lin(row) lin(col) !r spl(row) spl(col)` *terms in the model*

Replacing the quadratic terms with cubic spline effects makes very little difference in this example. The cubic spline model may fit slightly better on the basis of the average

SED (278.4 for quadratic model, 274.6 for the cubic spline model).

In general, the modeling approach for spatial analysis of variety trials should be done trying, as much as possible, to include terms in the model that are related to an identifiable source of variation. Furthermore, modeling small data sets such as TRIAL1.DAT, that has only 6 rows, should be done with caution in order to avoid overfitting. For small data sets to identify a parsimonious model is more important than to find a complex model. For this reason, for TRIAL1.DAT we have fitted a much simpler model than those showed before; this is model AR1xAR1+row (with random rows) that fits one more variance parameter (4) than the simple AR1xAR1. A random row term after the AR1xAR1 seems to provide with a variogram of residuals (Figure 4) similar to those previously found and a decrease in the Log-likelihood value with respect to the AR1xAR1 (Table 1). Thus, for TRIAL1.DAT with only 6 rows the only justifiable model to use seems to be AR1xAR1+row.

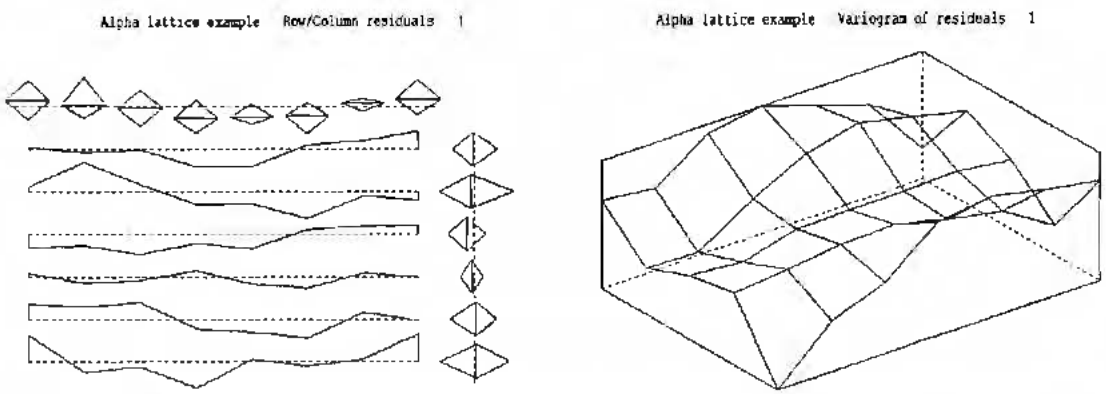

Figure 4. *Trellis plot (left) and variogram plot (right) for spatial models with random row term in the model*

Results of the different models, their Log-likelihood values, error variances and SED are given in Table 1.

Table 1. Results for various models. In the linear model random terms are bold

Linear model Mu + entry	**Error Variance Model**	**Log-likelihood**	**Error variance**	**Standard error of difference**
	IdxID	-232.13	424504	532.0
+ **rep**	IdxID	-223.48	206588	371.1
+ **rep** + **blk**	IdxID	-221.42	133293	326.5
	AR(1)xAR(1)	-221.47	413035	296.1
+ **row**	AR(1)xAR(1)	-220.07	171003	299.5
+ pol(row,-2)	AR(1)xAR(1)	-203.69	171946	292.3
+ lin(row) + **spl(row)**	AR(1)xAR(1)	-212.51	233964	293.6
+ lin(row) + **spl(row)**	IdxID	-213.91	148251	322.2
+ pol(row,-2) + pol(col,-2)	AR(1)xAR(1))	-189.61	126013	278.4
+ lin(row) + lin(col)	AR(1)xAR(1)	-211.56	405426	298.8
+ lin(row) + lin(col) + **spl(row)**	AR(1)xAR(1)	-208.90	172706	299.1
+ lin(row) + lin(col) + **spl(row)** + **spl(col)**	AR(1)xAR(1)	-206.84	110718	274.6
+ lin(row) + lin(col) + **spl(row)** + **spl(col)**	IdxID	-207.52	110699	290.1

Figure 5 compares the adjusted means of the RIBD, AR1 x AR1, AR1 x AR1 + pol(row,-2)+pol(col,-2), AR1 x AR1 + lin(row) + lin(col) + spl(row) + spl(col) and Ar1 x Ar1+row

models. The greatest differences are between the adjusted means of the RIBD and the other models. Notice that adding the extraneous model terms to the basic AR1 x AR1model has not greatly altered the adjusted means and that in this example the two-way spline model gives very similar adjusted means to the two-way quadratic polynomial model.

We conclude that there are highly significant row effects which are adequately fitted by including lin(row) !r spl(row) or possibly pol(row,2) in the model. Furthermore, there is a suggestion of column effects such that similar terms fitted to columns are significant ($P<0.05$) and so probably should be included but will have little effect on the adjusted means.

Once the row effects are in the model, the AR error correlations become non-significant and could be dropped. However, the estimated correlations are small and generally there is no loss in efficiency from leaving them in.

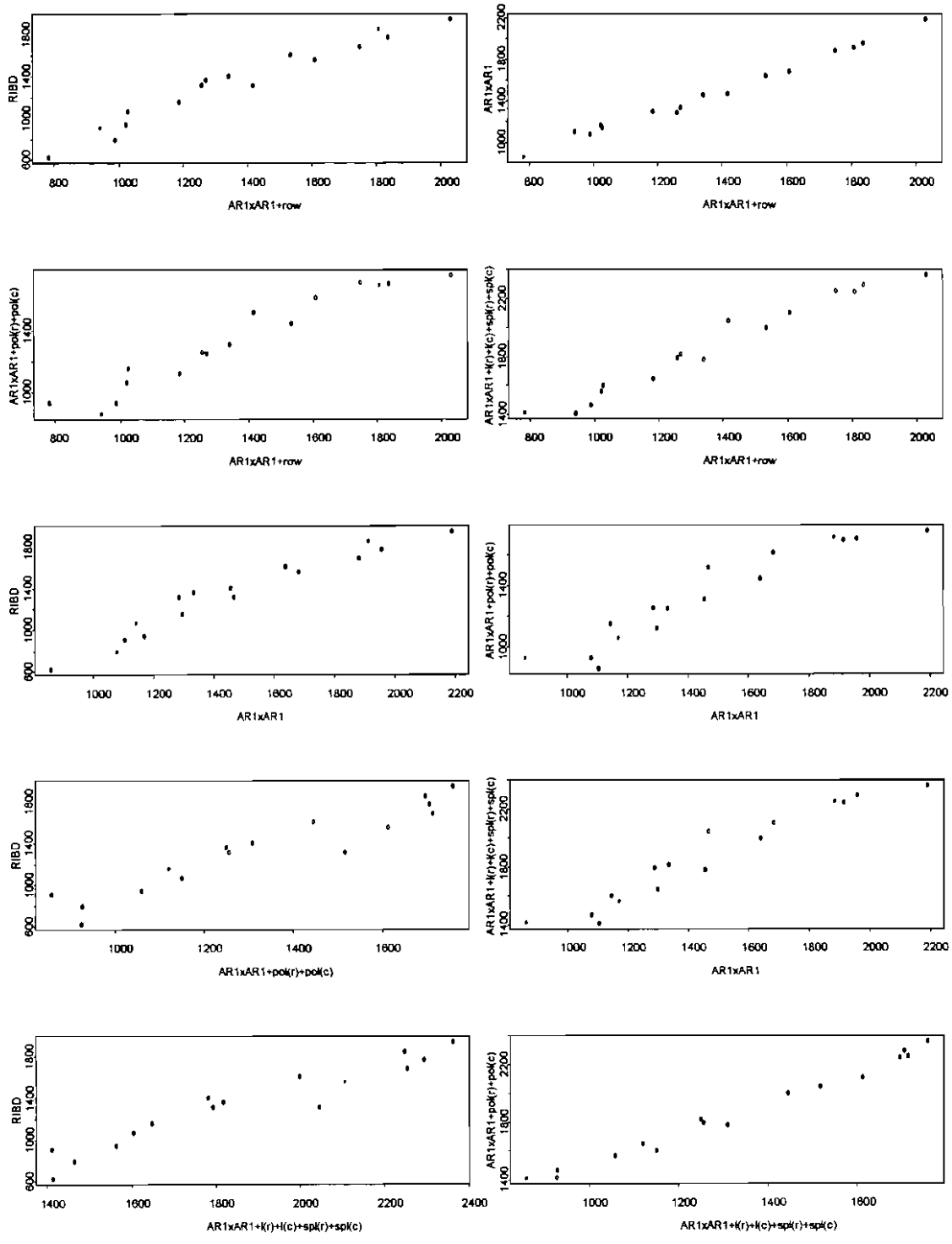

Figure 5. *Comparison of adjusted means from various models.*

6.1.2 TRIAL 2

I Preparing the data file

The second example is a variety trial designed as a row-column design (Nguyen and Williams, 1993) with 64 varieties planted in two contiguous replicates laid out in 8 rows and 16 columns. Data are in file TRIAL2.DAT.

First rows of the data file TRIAL2.DAT:

```
rep      row      col     variety yield(t/ha)
1        1        1        5       1.5318
1        1        2       19       2.2211
1        1        3       55       1.4589
1        1        4       23       1.2436
1        1        5       27       1.8989
1        1        6       38       1.3366
1        1        7       64       1.8966
...      ...      ...     ...      ...
```

II Analysis of the row-column design

The command file, RCD.AS, for analysis of the row-column design:

```
Row-column example
 rep 2
 row 8
 col 16
 variety 64
 y
TRIAL2.DAT !skip 1
y ~ mu variety !r rep rep.row rep.col
```

The linear model used in the file RCD.AS considers variety as fixed effect and rep, rep.row and rep.col as random effects.

Output file RCD.ASR for TRIAL2.DAT analyzed as row-column design:

```
ASREML [ 2 Sep 1999]  Row-colum example
28 Oct 1999 11:54:01.060    8.00 Mbyte    C:\RCD.AS
 Univariate analysis of y
Using      128 records [of      128 read from      128 lines of TRIAL2.DAT     ]
  Model term     Size Type     COL   Minimum     Mean      Maximum   #zero #miss
   1 rep            2 Factor     1         1    1.5000            2       0     0
   2 row            8 Factor     2         1    4.5000            8       0     0
   3 col           16 Factor     3         1    8.5000           16       0     0
   4 variety       64 Factor     4         1   32.5000           64       0     0
   5 y              1 Variate    5     1.151     2.539        4.783       0     0
   6 mu             1 Constant Term
   7 rep.row       16 Interaction  1 rep      :     2    2 row              :     8
   8 rep.col       32 Interaction  1 rep      :     2    3 col              :    16
 Forming 125 equations:  65 dense
 Initial updates will be shrunk by factor    0.548
 NOTICE:     34 (more) singularities,
 LogL=-31.9255    S2= 0.41906   64 df    0.1000   0.1000   0.1000    1.000
 LogL=-27.5664    S2= 0.26799   64 df    0.1795   0.3111   0.5202    1.000
 LogL=-25.8042    S2= 0.19528   64 df    0.3824   0.6574    1.085    1.000
 LogL=-25.5597    S2= 0.18114   64 df    0.6957   0.8706    1.225    1.000
 LogL=-25.5302    S2= 0.17985   64 df    0.9645   0.9148    1.218    1.000
 LogL=-25.5287    S2= 0.17960   64 df     1.060   0.9166    1.219    1.000
 Final parameter values                  1.0671  0.91656   1.2188   1.0000

 Source                 Model  terms     Gamma    Component    Comp/SE   % C
 rep                        2      2   1.06710     0.191650       0.55   0 P
 rep.row                   16      8  0.916555     0.164612       1.52   0 P
 rep.col                   32     32   1.21885     0.218903       2.57   0 P
 Variance                 128     64   1.00000     0.179598       3.78   0 P

 Analysis of Variance       DF     F-incr      F-adj  StndErrDiff
   6 mu                      1      52.00      20.89
   4 variety                63       0.84       0.84    0.5227

 Finished: 28 Oct 1999 11:54:04.960    LogL Converged
```

The Log likelihood for this model is –25.5287 with an error variance of 0.179598 and a SED of 0.5227.

III Spatial analysis using the AR1 x AR1

Following the approach of Gilmour et al. (1997) the error model AR1 x AR1 is fitted and the trends of the residuals are examined in the trellis plots and variogram. The command file for this analysis:

```
AR1 x AR1
 rep 2
 row 8
 col 16
 variety 64
 y
TRIAL2.DAT !skip 1
y ~ mu variety
1 2
row row AR1 0.1
col col AR1 0.1
```

The Log-likelihood value of the AR1 x AR1 model is –17.79 (Table 2) an increase of 7.74 with respect of the Log-likelihood of the row-column analysis despite the reduction in parameters. The SED, with respect to the row-column analysis decreases from 0.523 to 0.443. The trellis plot of residual and variogram for the AR1 x AR1 analysis are in Figure 6. The smoothness of the variogram is indicative of strong non-stationary trends. The shape is indicative of a curved pattern for rows and a flatter pattern for columns. The trellis plot shows that both appear curvilinear.

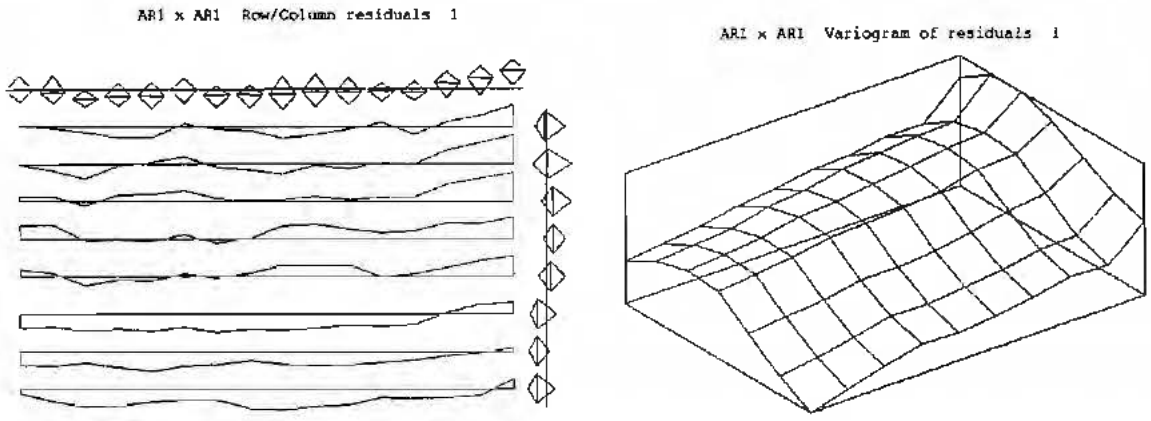

Figure 6. *Trellis plot (left) and variogram (right) of residuals from the AR1xAR1 model. Row distance is on the left axis, column distance is on the right axis.*

IV Spatial analysis using the AR1 x AR1 plus extensions

Our next model fits cubic smoothing splines to both rows and columns. Note that each spline, as implemented in ASREML requires two terms to be fitted: lin(row) and lin(col) as fixed terms, spl(row) and spl(col) as random terms. The spl() terms may be omitted to just fit a linear trend. Since this involves adding fixed terms, we test the linear terms using F-ratios and then the spl() terms using likelihood changes (Table 2). We see a huge increase in the REML logl values going from the lin(r) + lin(c) model (REML logl=-19.88) to the lin(r) + lin(c) + spl(row) + spl(col) model (REML logl=-9.24) and the F-ratios for lin(row) and lin(col) in this final model are huge (70.4 and 41.6 respectively). The trellis plot and variogram from this spline model is acceptable (Figure 7).

Output of the AR1 x AR1 +lin(r) + lin(c)+spl(r) + spl(c):

```
Forming  87  equations:  87  dense
 Initial updates will be shrunk by factor     0.548
 NOTICE:     1 (more) singularities,
 LogL=-9.23934  S2= 0.14822  62 df  1.302  0.1182  1.000  0.9413E-02  0.1151
 LogL=-9.23934  S2= 0.14822  62 df  1.302  0.1182  1.000  0.9430E-02  0.1152
 Final parameter values                1.3016 0.11816 1.0000 0.94340E-02 0.11516

 Source                Model  terms      Gamma      Component    Comp/SE   % C
 spl(row)                  6      6    1.30162       0.192924       1.10   0 P
 spl(col)                 14     14   0.118156      0.175129E-01    0.99   0 P
 Variance                128     62    1.00000       0.148219       4.75   0 P
 Residual          AR=AutoR      8  0.943403E-02   0.943403E-02    0.04   0 U
 Residual          AR=AutoR     16   0.115165       0.115165       0.51   0 U

 Analysis of Variance       DF     F-incr      F-adj  StndErrDiff
  10 spl(col)               14       8.85       3.72
   9 spl(row)                6      14.20       8.27
   6 mu                      1    4409.93       6.80
   4 variety                63       1.86       1.00    0.4038
   7 lin(row)                1      70.96      70.25
   8 lin(col)                1      41.60      41.60

Finished: 28 Oct 1999 14:18:35.220   LogL Converged
```

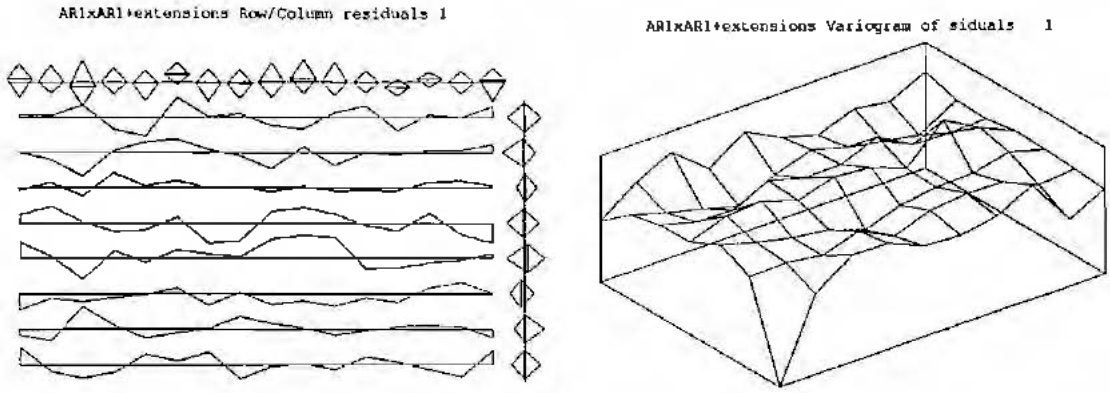

Figure 7. *Trellis plot (left) and variogram (right) of residuals from model with lineal plus spline effects in row and column. Row distance is on the left axis, column distance is on the right axis.*

Table 2. Results for various models. In the linear model, random terms are bold.

Linear model Mu + variety	**Error Variance Model**	**REML Log-Likelihood**	**Error Variance**	**Standard Error of Difference**
+ **rep** + **row** + **column**	IdxID	-25.53	0.18	0.523
	AR(1)xAR(1)	-17.79	0.62	0.443
+ lin(r) + lin(c)	AR(1)xAR(1)	-19.88	0.46	0.450
+ lin(r) + lin(c) + **spl(r)**	AR(1)xAR(1)	-17.45	0.30	0.441
+ lin(r) + lin(c) + **spl(r)** + **spl(c)**	AR(1)xAR(1)	-9.24	0.15	0.404

6.2 SPATIAL ANALYSIS OF A VARIETY TRIAL WITHOUT REPLICATES

I Preparing the data file

A total of 280 varieties were planted in a rectangular array of 14 rows and 31 columns. There is one check variety planted in every 3rd column (1, 4, ..., 31). The data file name is UNREP.DAT and the first rows of the data file are shown below. Columns are: check (1 = check entry, 2 = no check entry), entry, column, row and yld.

First rows of the data file UNREP.DAT:

check	entry	col	row	Yld
1	0	1	1	7.21504
2	1	2	1	5.13106
2	2	3	1	5.97648
1	0	4	1	5.42136
2	3	5	1	4.86704
2	4	6	1	7.70989
1	0	7	1	6.10659
2	5	8	1	7.87475
2	6	9	1	7.89012
1	0	10	1	6.61043
2	7	11	1	6.27366
2	8	12	1	5.44724

II Spatial analysis using the AR1 x AR1 model

The following lines constitute the ASREML command file for the initial spatial analysis.

Command file UNREP.AS for analyzing UNREP.DAT using the AR1 x AR1 model:

```
Unreplicated analysis with ASREML
 check    2
 entry  281 !I
 col     31
 row     14
 yld
unrep.dat !skip 1
yld ~ mu !r entry
1 2
row row AR1 0.1
col col AR1 0.1
```

In this model, the variety including check is considered a random effect. Some workers (for example Cullis et al., 1989) fit the difference between CHECK lines and TEST lines as a fixed effect. However, since the check line is so highly replicated, it is easier and almost equivalent to do as we have here. The two-dimensional AR1 x AR1 model for the residuals is specified in the last three lines.

Output file UNREP.ASR for UNREP.DAT analyzed using the AR1 x AR1 model:

```
ASREML [ 9 Aug 1999]  Spatial analysis of an unreplicated field trial
 11 Aug 1999 10:27:19.230  32.00 Mbyte    unrep
  Reading unrep.dat  FREE FORMAT skipping  1  line
 Univariate analysis of yld
 Using      434 records [of      434 read from      434 lines of unrep.dat]
  Model term      Size Type    COL   Minimum    Mean    Maximum   #zero #miss
   1 check          2 Factor    1         1    1.6452         2         0     0
   2 entry        281 Factor    2         1   91.6452       281         0     0
   3 col           31 Factor    3         1   16.0000        31         0     0
   4 row           14 Factor    4         1    7.5000        14         0     0
   5 yld            1 Variate   5     2.115     6.340     9.333         0     0
   6 mu             1 Constant Term
    14  AR=AutoR    0.10
    31  AR=AutoR    0.10
  Forming  283  equations:  2  dense
 Initial updates will be shrunk by factor    0.548
 LogL=-237.568   S2=  1.0348   433 df  0.1000   1.000   0.1000       0.1000
 LogL=-225.231   S2= 0.81449   433 df  0.4252   1.000   0.8531E-01   0.2133E-
01
 LogL=-213.282   S2= 0.54989   433 df   1.405   1.000   0.9564E-01  -0.1001
 LogL=-211.827   S2= 0.50149   433 df   1.783   1.000   0.1068      -0.2064
 LogL=-211.731   S2= 0.49989   433 df   1.815   1.000   0.1186      -0.2359
 LogL=-211.725   S2= 0.49944   433 df   1.824   1.000   0.1234      -0.2414
 Final parameter values                1.8262   1.000   0.12493     -.24263

 Source                Model  terms     Gamma     Component    Comp/SE   % C
 entry                   281    281   1.82619      0.912075       7.18   0 P
 Variance                434    433   1.00000      0.499443       8.91   0 P
 Residual             AR=AutoR   14  0.124930      0.124930       1.77   1 U
 Residual             AR=AutoR   31 -0.242635     -0.242635      -2.65   0 U

 Analysis of Variance       DF     F-incr      F-adj  StndErrDiff
   6 mu                      1    8329.18    8329.18

                     Solution       Standard Error    T-value     T-prev
   6 mu
                  1    6.44048       0.705695E-01      91.26
   2 entry                281 effects fitted
 Finished: 11 Aug 1999 10:27:23.570  LogL Converged
```

A variogram of the residuals and trellis plots are shown in Figure 8. As in the case of the replicated trial, this AR1xAR1 model can be improved.

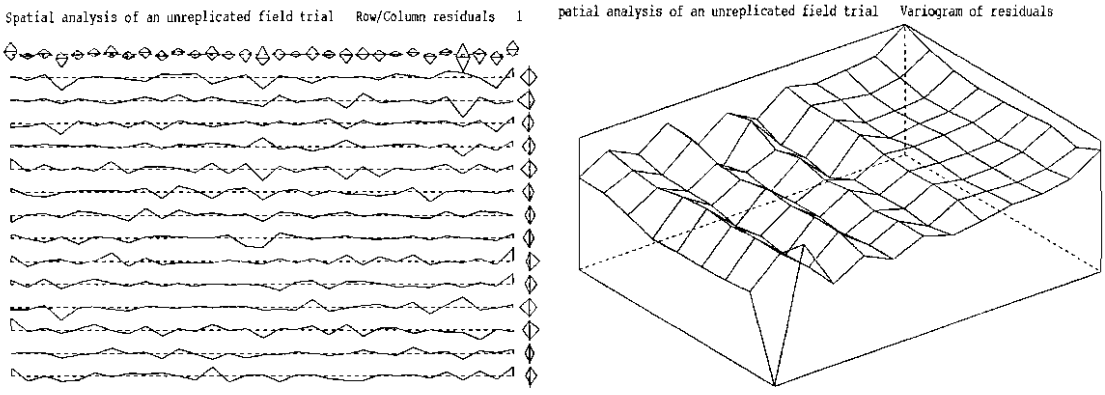

Figure 8. *Trellis plot (left) and variogram (right) of residuals for the unreplicated data of the model AR1xAR1. Row distance is on the left axis, column distance is on the right axis.*

Looking at the variogram, we see strong ridges associated with columns but they do not persist across the whole variogram. This indicates column effects which locally have a sawtooth pattern but not consistently. This is also indicated by the negative autocorrelation parameter associated with columns. Looking at the trellis plot (Figure 8), we can see the sawtooth pattern does seem to persist over all the columns. The columns containing the checks (every fourth) also stand out because they have bigger residuals (having true replication). We therefore add random column effects and fixed sawtooth effect to the model by changing the model line to read

```
yld ~ mu altcol !r entry  col
```

where **altcol** is a new covariate added to the data which is 1 for odd columns and 0 for even columns. Some models are summarised in table 3.

Adding *col* to the initial model increased the LogLikelihood to -201.4, a significant increase of 10.3. Adding altcol to the base model, altcol was significant with an F ratio

of 38.3. Adding *col* to this model still improved the likelihood significantly but altcol remains significant.

Table 3 Summary of models fitted to unreplicated trial. All models fitted with AR1xAR1 error correlation. Random effects are in bold.

Model	REML loglikelihood
mu + **entry**	-211.7
+ **col**	-201.4
+ altcol	-199.5
+ altcol + **col**	-196.6

This analysis assumes that entries are truly randomly distributed. In experiments where entries come from families and are laid out in the field in family blocks, then these column effects could be family effects. Thus, appropriate design is required for spatial analysis.

Figure 9 show the variogram and residual plot after adjusting for AR1xAR1 plus column effects. The variogram is fine.

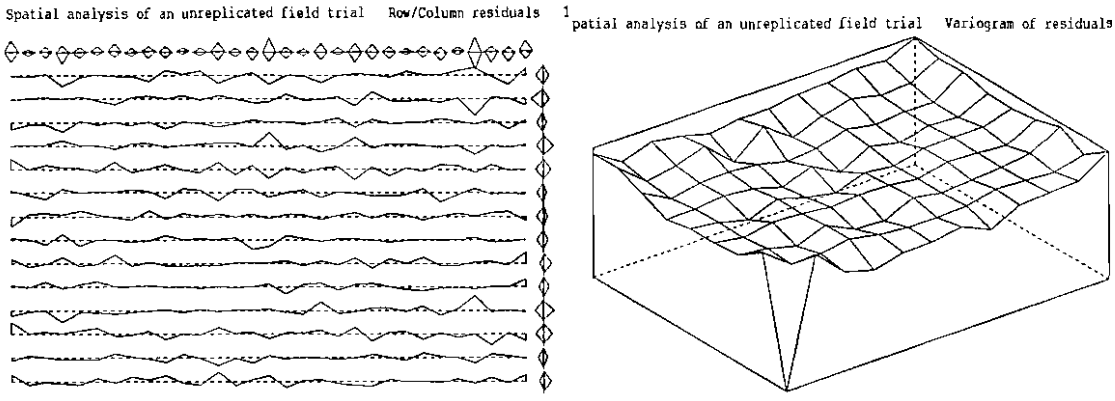

Figure 9. *Trellis plot (left) and variogram (right) of residuals for the unreplicated data corresponding to the analysis of the model AR1xAR1 + col. Rows are on the right side and columns are on the left side.*

This is a reasonable model for this data. We note that that the residuals tend to be larger in the CHECK columns (1, 4, ...31). This is because these are the replicated plots and so there is a good estimate of the mean, hence of the residual. Second, we note that column effects which appear present would be better estimated if check plots occurred in every column. For this reason, check plots are often run diagonally. These are design issues which need to be addressed.

Predicted values for checks and entries of UNREP.DAT as listed in UNREP.SLN for the model with only col :

The file UNREP.SLN contains the Best Linear Unbiased Predictors (BLUPs) for the entries including the CHECK line. The predicted mean is obtained by adding the mean (mu = 6.450) to the BLUPs. Thus for the check line (= Entry 0), the mean is 6.1368 = 6.450 - 0.3132. The adjusted mean for entry 1 is 5.7349 compared with the unadjusted data value of 5.131.

The mean appears at the beginning of the output followed by the yield effect of each entry, the column effects and lastly the 14 x 31 residuals. The last column contains the standard error of the effects and, for the residuals, is the predicted value for that plot. In this case the predicted value involves the mean, entry effect and the column effect. Thus, adding the column 1 effect (0.3281) to the CHECK mean (6.1368) gives 6.465.

mu		1	6.450	0.9592E-01
entry		0	-0.3132	0.1436
entry		1	-0.7151	0.5426
entry		2	-0.4548	0.5426
entry		3	-1.077	0.5426
...		...	...	...
...		...	...	...
entry		287	0.3728	0.5426
entry		288	1.756	0.5426
entry		289	-0.5668	0.5426
entry		290	-2.481	0.5426
col		1	0.3281	0.1610
col		2	-0.1857	0.2230
col		3	0.6567E-01	0.2230
col		4	-0.4431	0.1609
...		...	...	...
...		...	...	...
col		26	-0.3041	0.2230
col		27	0.3395	0.2230
col		28	-0.2862	0.1609
col		29	0.6290E-01	0.2230
col		30	-0.3247	0.2230
col		31	0.5475	0.1610
	Residual	1	0.7503	6.465
	Residual	2	-0.4180	5.549
	Residual	3	-0.8429E-01	6.061
	Residual	4	-0.2722	5.694
...	...	...	...	...

7 References

Brownie, C., D.T. Bowman, and J.W. Burton. 1993. Estimating spatial variation in analysis of data from yield trials: A comparison of methods. Agronomy Journal, 85:1244-1253.

Cullis, B.R., and A.C. Gleeson. 1991. Spatial analysis of field experiments-an extension to two dimensions. Biometrics, 47: 1449-1460.

Cullis, B.R., Lill, W. J., Fisher, J. A., Read, B. J., and Gleeson, A.C. 1989. A new procedure for the analysis of early generation variety trials. Applied Statistics, 38: 361-375.

Cullis, B.R., Bev Gogel, A.P. Verbyla and Robin Thompson 1998 Spatial Analysis of Multi-environment early generation variety trials. *Biometrics* **54**: 1-18.

Gleeson, A.C., and B.R. Cullis. 1987. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. Biometrics, 43:277-288.

Gilmour, A. R., B.R. Cullis, and A.P. Verbyla. 1997. Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. Journal of Agricultural, Biological, and Environmental Statistics, Vol. 2, 269-293.

Gilmour, A. R., B.R. Cullis, S.J. Welham, and R. Thompson. 1999. ASREML Reference Manual, Biometrics Bulletin 3, NSW.Agriculture, ORANGE, 2800, Australia.

Grondona, M.O., J. Crossa, P.N. Fox, and W.H. Pfeiffer. 1996. Analysis of variety trials using two-dimensional separable ARIMA processes. Biometrics, 52:763-770.

Nguyen, N-K, and Williams, E.R. (1993). An algorithm for constructing optimal resolvable row-column designs. Austral. J. Statis. 35: 363-370.

Patterson, H.D. and Williams, E.R. (1976). A new class of resolvable incomplete block designs. Biometrika, 63:83-92.

Biplots of Linear-Bilinear Models for Studying Crossover Genotype × Environment Interaction

Jose Crossa,* Paul L. Cornelius, and Weikai Yan

ABSTRACT

Linear-bilinear models, such as the Shifted Multiplicative Model (SHMM) and Sites Regression Model (SREG), have been used to develop clustering procedures for finding subsets of sites (or cultivars) without cultivar crossover interaction (non-COI). Biplots of these models are useful for visual evaluation of cultivar responses across environments. The main purposes of this study were to investigate (i) $SREG_2$ and $SHMM_2$ biplots with the first multiplicative components constrained to be non-COI $SREG_1$ and $SHMM_1$ solutions, (ii) how the biplots can be used for identifying subsets of sites and cultivars with different levels of COI and with non-COI, and (iii) how these biplots compare with results obtained when clustering only sites or cultivars without cultivar rank change. Transformed and untransformed data from two multienvironment cultivar trials were used for illustration. Biplots from $SHMM_2$ and $SREG_2$ models graphically display the interaction variation due to low level COI or non-COI (first multiplicative term) versus the interaction variation due to COI (second multiplicative term). The biplots obtained by means of the non-COI first term constrained solution of the $SREG_2$ and $SHMM_2$ models have the same interpretability properties as the standard biplots obtained by means of the unconstrained solution. With the unconstrained and constrained solutions, it is possible to identify subsets of sites and cultivars with low level COI and non-COI. Biplots based on unscaled or scaled data produced similar results. Groups of sites and cultivars with low level COI and non-COI were similar to those found when only sites (or cultivars) were clustered into non-COI groups using the SHMM and SREG clustering approach.

MULTIPLICATIVE MODELS for multisite cultivar trials have been used for studying genotype × environment interaction (GEI) and for developing methods for clustering sites or cultivars into groups with statistically negligible crossover interaction (COI) (Cornelius et al., 1992, 1993; Crossa et al., 1993, 1995, 1996; Crossa and Cornelius, 1993, 1997; Abdalla et al., 1997). Multiplicative models have an additive (linear) component (i.e., intercept, main effects of sites and/or cultivars) and a multiplicative (bilinear) component (GEI) and thus are also named linear-bilinear models (Cornelius and Seyedsadr, 1997). Two types of linear-bilinear models are suitable for grouping sites and cultivars without cultivar rank change: the shifted multiplicative model (SHMM), $\bar{y}_{ij} = \beta + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$, and the sites regression model (SREG), $y_{ij} = \mu_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}$, where $\bar{y}_{ij}$ is the mean of the ith cultivar in the jth environment for g cultivars and e sites ($i = 1,2,...,g$ and $j = 1,2...e$); β is the shift parameter; μ_j is the site mean; $\lambda_k (\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_t)$ are scaling constants (singular values) that allow the imposition of orthonormality constraints on the singular vectors for cultivars, $\alpha_k = (\alpha_{1k},...,\alpha_{gk})$, and sites, $\gamma_k = (\gamma_{1k},...,\gamma_{ek})$, such that $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$ and $\sum_i \alpha_{ik}\alpha_{ik'} = \sum_j \gamma_{jk}\gamma_{jk'} = 0$ for $k \neq k'$; α_{ik} and γ_{jk}, for $k = 1,2,3,...$, are called *primary*, *secondary*, *tertiary*,..., effects of ith cultivar and the jth site, respectively; ε_{ij} is the residual error assumed to be NID $(0, \sigma^2/r)$ (where σ^2 is the pooled error variance and r is the number of replicates). The number of bilinear terms $t \leq \min(g, e)$. Estimates of the multiplicative parameters in the kth bilinear term are obtained as the kth component of the singular value decomposition (SVD) of the deviations from the additive part of the model. In the SHMM model, the bilinear terms absorb the environmental and genotypic main effects and the GEI, whereas in the SREG model, only the main effects of cultivars plus the GEI are absorbed into the bilinear terms.

If SHMM and SREG models with one multiplicative component ($SHMM_1$ and $SREG_1$) are adequate for fitting the data (second, third, and higher order multiplicative components are negligible) and primary effects of the sites, $\hat{\gamma}_{j1}$, are either all non-positive or all non-negative, $SHMM_1$ and $SREG_1$ predict non-COI. On the contrary, if $\hat{\gamma}_{j1}$ are of different signs, $SHMM_1$ and $SREG_1$ models predict COI. Moreover, the non-COI property of $SHMM_1$ and $SREG_1$ (when $\hat{\gamma}_{j1}$ are either all non-positive or all non-negative) is a consequence of a proportionality condition, i.e., cultivar differences in any one site are proportional to their differences in any other site.

In various clustering studies based on SHMM or SREG (Cornelius et al., 1992, 1993; Crossa and Cornelius, 1997), the measure of distance (i.e., dissimilarity) between a pair of sites was the residual sum of squares (RSS) after fitting $SHMM_1$ or $SREG_1$, $RSS(SHMM_1)$ or $RSS(SREG_1)$, respectively. Seyedsadr and Cornelius (1993) proved that if $e \leq g$, $RSS(SHMM_{e-1}) = RSS(SREG_{e-1})$. Thus, for a subset of data containing only two sites, $RSS(SHMM_1) = RSS(SREG_1)$. If the resulting $\hat{\gamma}_{j1}$ have the same sign, $RSS(SHMM_1)$ is a non-COI solution; but if $\hat{\gamma}_{j1}$ are of different signs, constrained $SHMM_1$ and $SREG_1$ solutions need to be computed. Crossa et al. (1993), in clustering sites into groups with non-COI, used constrained least squares (LS) $SHMM_1$ solutions for pairs of sites needing constrained solutions, but Cornelius et al. (1993), in clustering cultivars into groups with non-COI, used constrained singular value

J. Crossa, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Lisbon 27, Apdo. Postal 6-641, 06600 Mexico D.F., Mexico; P.L. Cornelius, Dep. of Agronomy and Dep. of Statistics, Univ. of Kentucky, Lexington, KY 40546-0091; W. Yan, Dep. of Plant Agriculture, Univ. of Guelph, Guelph, ON, Canada N1G2W1. The investigation reported in this paper (No. 00-06-100) relates to a project of the Kentucky Agric. Exp. Stn. and is published with approval of the Director. Received 9 May 2001. *Corresponding author (j.crossa@cgiar.org).

Published in Crop Sci. 42:619–633 (2002).

Abbreviations: SHMM, Shifted Multiplicative Model; SREG, Site Regressions Model; GEI, genotype × environment interaction; COI, crossover interaction; LS, least squares; SVD, singular value decomposition.

Table 1. Grain-yield rank of nine (G1–G9) maize cultivars at 20 test sites (Trial 1) and the standard error of cultivar means (SE) (kg ha^{-1}) at each site.

	Site																			
Cultivar	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
G1	4	8	2	7	7	6	8	1	7	2	7	8	5	7	5	8	7	8	9	7
G2	6	7	7	6	5	7	9	2	6	4	8	6	7	5	9	6	8	6	7	8
G3	5	5	6	5	9	5	4	3	4	7	6	4	6	6	6	4	6	7	3	3
G4	3	3	8	1	1	4	7	9	5	6	1	1	1	1	1	1	4	1	8	4
G5	8	6	1	3	3	2	2	8	2	9	4	2	3	3	2	2	2	5	2	1
G6	2	1	5	2	2	1	1	7	1	3	3	5	2	2	3	3	1	2	4	2
G7	7	2	3	9	8	3	3	5	8	5	5	7	8	8	7	7	9	3	1	9
G8	1	9	4	8	6	9	5	4	9	1	9	9	9	9	8	9	3	9	6	6
G9	9	4	9	4	4	8	6	6	3	8	2	3	4	4	4	5	5	4	5	5
SE	480	274	538	251	324	328	317	425	682	142	487	386	429	550	361	270	371	567	722	296

decompositions (SVD) to obtain $SHMM_1$ solutions. The constrained SVD solution will force only the most extreme primary effect of a site (located at left or right of the graph) to be zero, whereas the constrained LS solution will assign a value of zero to primary effects of as many sites as necessary to assure that site primary effects are either all non-negative or all non-positive.

Biplots are useful for summarizing and approximating patterns of response that exist in the original data (Gabriel, 1971, 1978). Yan et al. (2000) presented standard biplots of the SREG model that helped enhance its interpretation for selecting the best performing cultivars in subsets of sites. The authors proposed (i) connecting the markers of the farthest (most responsive) cultivars in the biplot such that they are the corners (vertices) of an irregular polygon and (ii) for each side of the polygon, drawing a line segment perpendicular to that side and passing through the origin. These line segments subdivide the polygon into sectors involving different subsets of sites and cultivars. The cultivar that is at the polygon corner located in one sector is the best performer (due to large positive GEI) in sites with markers included in that sector, but it is the worst performer (due to large negative GEI) in sites with markers located in the opposite sector of the biplot. The biplot from the SREG model shows that ideal cultivars should have large primary effects (high mean yield) and near-zero secondary effects (more stable) and the ideal sites should have large primary effects (high power to discriminate cultivars) and small secondary effects. Such properties tend to occur if the primary effects of cultivars are highly correlated with the cultivar means (Yan et al., 2001).

For $SHMM_2$ and $SREG_2$ models, the biplot of the first two multiplicative components would represent the graph of the interaction variation due to non-COI (first multiplicative term) (or proportionality of cultivar response in sites) versus the interaction variation due to COI (second multiplicative term) (or disproportionality of cultivar response in sites). This is accomplished if, and only if, the scores of the first singular vector for sites, $\hat{\gamma}_{j1}$, are all of the same sign. If $\hat{\gamma}_{j1}$ are of different signs, a constrained solution for $SHMM_2$ and $SREG_2$ is required, such that the first multiplicative term should show a non-COI pattern. For $SHMM_2$, this is simply obtained by constraining the first multiplicative term by the standard constrained SVD solution and using the second multiplicative component of its SVD as the second multiplicative term. For the SREG model, the SVD non-COI solution is not that simple.

Previous research using the SHMM and SREG models led to the development of clustering procedures for finding subsets of sites with non-COI or subsets of cultivars with non-COI (Cornelius et al., 1992, 1993; Crossa et al., 1993, 1995, 1996; Crossa and Cornelius, 1993, 1997). However, these procedures do not simultaneously identify non-COI subsets of cultivar and sites. The main purposes of this study were to investigate (i) $SREG_2$ and $SHMM_2$ biplots with the first multiplicative components constrained to be non-COI $SREG_1$ and $SHMM_1$ solutions and to compare these with the biplot of Yan et al. (2000) in cases where the unconstrained solution does not yield a non-COI solution, (ii) how the biplots can be used for identifying subsets of sites and cultivars with different levels of COI and with non-COI, and (iii) how these biplots compare with results obtained when clustering only sites or cultivars without cultivar rank change.

MATERIALS AND METHODS

Experimental Data

Trial 1

The data are the same as in Cornelius et al. (1992, Table 5), Cornelius et al. (1996, Table 2), and Crossa and Cornelius (1993, Fig. 75-1) where nine ($g = 9$) CIMMYT maize (*Zea mays* L.) cultivars were evaluated in a randomized complete block design with four ($r = 4$) replications at each of the 20 ($e = 20$) international sites. The rank of the cultivars in each site and the standard error of cultivar means at each site are shown in Table 1.

Trial 2

This data set involves 11 winter wheat (*Triticum aestivum* L.) cultivars tested in 26 environments (year–site combinations during 1996–1998), extracted from the Ontario winter wheat performance trial database maintained at the Univ. of Guelph. The Ontario winter wheat performance trials are sponsored by the Ontario Ministry of Agriculture, Food and Rural Affairs, the Ontario Wheat Producers Marketing Board, and sponsors of the varieties to provide information to Ontario wheat growers about the performance of the available winter wheat cultivars. The trials at the various test sites were in randomized complete block designs with four to six replicates, but only the cultivar × environment mean yield data of $g = 11$ and $e = 26$ are available. The ranks of the cultivars in each test environment are presented in Table 2.

Table 2. Grain-yield ranks of 11 (G1–G11) winter wheat cultivars (Trial 2) at 26 test sites.

	Cultivar										
Site	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11
1	3	11	5	6	8	10	4	2	7	1	9
2	4	11	5	10	1	9	7	6	2	3	8
3	1	9	2	8	11	10	5	3	6	6	4
4	10	5	6	9	11	4	3	8	1	2	7
5	5	10	1	11	7	6	3	4	9	2	8
6	7	4	8	11	10	5	6	9	1	2	3
7	6	1	8	3	9	11	5	4	2	10	7
8	6	7	1	5	10	11	9	3	4	2	8
9	4	11	2	1	3	9	10	5	6	7	8
10	10	3	2	8	11	5	4	6	1	8	7
11	3	1	7	11	10	5	4	8	2	9	6
12	7	10	1	5	11	8	6	4	3	2	9
13	2	4	9	6	4	11	3	1	8	10	7
14	6	3	7	10	11	7	2	5	1	9	4
15	6	3	5	11	2	9	9	3	7	8	1
16	8	3	6	7	11	2	5	4	9	10	1
17	2	2	10	11	9	8	5	6	1	7	4
18	9	5	1	9	3	7	11	6	2	4	8
19	8	11	1	9	10	4	3	2	6	7	5
20	10	1	3	4	7	8	6	9	1	11	5
21	3	4	5	7	9	10	11	8	1	6	2
22	2	6	7	5	10	9	4	11	1	8	2
23	6	4	6	3	8	2	6	11	1	10	8
24	9	11	2	6	6	10	4	3	8	1	5
25	6	10	4	5	11	9	3	1	8	7	2
26	6	4	5	10	11	7	2	8	3	9	1

Scaled Data and Notation

In both trials, the non-COI unconstrained and constrained SVD and LS SREG and SHMM solutions were computed on the basis of the unscaled values of the cell means ($\bar{y}_{ij}$) and on the basis of the scaled data ($\bar{z}_{ij}$). The scaled data were computed as in Crossa and Cornelius (1997), that is, $\bar{z}_{ij} = \bar{y}_{ij}/(\sqrt{s_j^2/r})$, where s_j^2 is the error variance in the jth site and r is the number of replicates.

The notation used for the $SREG_2$ and $SHMM_2$ analyses and the corresponding biplots using unscaled and scaled data and unconstrained and constrained solutions is as follows: the first capital letter in parenthesis denotes the type of data used (U = unscaled, S = scaled) and the second capital letter represents the type of solution applied (U = unconstrained; C_{SVD} = constrained SVD; C_{LS+1} = constrained LS $SREG_1$ as the first term and with the second term taken as the first component of the SVD of residuals from the $SREG_1$ constrained LS solution). Thus, $SREG_2$(U/U) denotes the sites regression model on unscaled data and applying an unconstrained solution; $SREG_2$(U/C_{SVD}) denotes the sites regression model on unscaled data and applying a constrained SVD solution; $SREG_2$(S/U) denotes the sites regression model on scaled data and using an unconstrained solution; $SREG_2$(S/C_{SVD}) denotes the sites regression model on scaled data and using a constrained SVD solution; $SREG_2$(S/C_{LS+1}) denotes the sites regression model on scaled data and using a constrained LS solution. Similar notation is used for biplots of the $SHMM_2$ model where only constrained SVD solutions are computed.

Constrained SVD Non-COI Solution for the SHMM Model

The constrained SVD non-COI solution of the SHMM model for two or more cultivars is described in Cornelius et al. (1993), and for two or more sites, they are given in Crossa et al. (1995, 1996). Briefly, the matrix subjected to SVD is $\mathbf{Z} = \{z_{ij}\} = \{\bar{y}_{ij} - \hat{\beta}\}$. The residual sum of squares for $SHMM_1$ is $RSS(SHMM_1) = \text{trace}(\mathbf{Z}'\mathbf{Z}) - L_1$, where L_1 is the largest eigenvalue of $\mathbf{Z}'\mathbf{Z}$. In this case, the value of $\hat{\beta}$ is selected such that the smallest or the largest $\hat{\gamma}_{j1}$ is zero. For a pair of sites, 1 and 2, and g cultivars, the closed-form SVD non-COI $SHMM_1$ solutions for $\hat{\beta}$ are given by

$$\hat{\beta} = 0.5\left\{(\bar{y}_{.1} + \bar{y}_{.2}) \pm \left[(\bar{y}_{.1} + \bar{y}_{.2})^2 - \frac{4\sum_i \bar{y}_{i1}\bar{y}_{i2}}{g}\right]^{1/2}\right\},$$

provided that

$$(\bar{y}_{.1} + \bar{y}_{.2})^2 - \frac{4\sum_i \bar{y}_{i1}\bar{y}_{i2}}{g} > 0.$$

Both solutions force $\mathbf{Z}'\mathbf{Z}$ to be diagonal. One solution has $\hat{\gamma}_{i1} = 0$ and $\hat{\gamma}_{i2} = 1$ and the other $\hat{\gamma}_{i1} = 1$ and $\hat{\gamma}_{i2} = 0$. For either value of $\hat{\beta}$, $RSS(SHMM_1) = \text{trace}(\mathbf{Z}'\mathbf{Z}) - L_1$ reduces to the sum of squared deviations of the cultivar means in the site with $\hat{\delta}_{j1}$ so that the solution is the one that gives the minimum $RSS(SHMM_1) = \min[\sum_i(\bar{y}_{i1} - \hat{\beta})^2, \sum_i(\bar{y}_{i2} - \hat{\beta})^2]$.

For more than two sites, the SVD non-COI $SHMM_1$ solution does not exist in closed form (Cornelius et al., 1993). If site m is selected to have $\hat{\gamma}_{m1} = 0$, the constraint is $\hat{\lambda}_1\hat{\gamma}_{m1} = \sum_i\hat{\alpha}_{i1}(\bar{y}_{im} - \hat{\beta}) = 0$ for which

$$\hat{\beta} = \frac{\sum_i \hat{\alpha}_{i1}\bar{y}_{im}}{\sum_i \hat{\alpha}_{i1}},$$

which is iteratively computed and where α_{i1} is estimated by the SVD of $\mathbf{Z}$. (Note that here $\mathbf{Z}$ also changes iteratively.)

Constrained SVD Non-COI Solution for the SREG Model

For the constrained SVD non-COI solution for the SREG model, a solution is required such that $\mathbf{Z} = \{z_{ij}\} = \{\bar{y}_{ij} - \hat{\mu}_j\}$ has elements of its first right singular vector all of the same sign (or zero). The proposed solution to this problem is to put $\hat{\mu}_j = \bar{y}_{.j} + \hat{\beta}$ and choose $\hat{\beta}$ to satisfy the required condition. Note that, after shifting the $\hat{\mu}_j$ values (from $\bar{y}_{.j}$), the $\hat{\mu}_j$ should no longer be perceived as estimates of site means.

If a constrained solution is needed, the $\hat{\gamma}_{j1}$ values in the unconstrained solution will contain both positive and negative values. Let S^- and S^+ denote the sum of squares of the negative and positive $\hat{\gamma}_{j1}$ values, respectively. If $S^- < S^+$, choose the

site with the most negative $\hat{\gamma}_{jl}$ value to be the site to have its $\hat{\gamma}_{jl}$ value forced to zero in the constrained solution. Conversely, *if* $S^- > S^+$, choose the site with the largest positive $\hat{\gamma}_{jl}$ value to have its $\hat{\gamma}_{jl}$ value forced to zero. Suppose these rules lead to a Site m as the site so chosen, i.e., $\hat{\beta}$ will be chosen to force $\hat{\delta}_{ml} = 0$. Solutions for $\hat{\beta}$ are

$$\hat{\beta} = \pm \sqrt{\frac{-\sum_i [\sum_{j \neq m} \hat{\gamma}_{jl}(\bar{y}_{ij.} - \bar{y}_{j.})](\bar{y}_{im.} - \bar{y}_{.m.})}{g \sum_{j \neq m} \hat{\gamma}_{jl}}} \qquad [1]$$

Derivation of this result is given in Appendix 1. Iterate until the value of $\hat{\beta}$ converges, consistently choosing either the positive or negative solution on every iteration. At convergence, the quantity under the radical in Eq. [1] is necessarily positive. In practice, to ensure that the iteration actually gets started, replace the quantity under the radical with its absolute value if it is negative.

Typically, the negative solution for $\hat{\beta}$ will make most of the $\hat{\alpha}_{il}$ values of the same sign as the nonzero $\hat{\gamma}_{jl}$ values. The positive solution for $\hat{\beta}$ will have the opposite effect. Absolute values of the $\hat{\gamma}_{jl}$, and also of the $\hat{\gamma}_{jl}$, will be the same for either solution, but this will not hold for the $\hat{\alpha}_{il}$ or for the $\hat{\alpha}_{i2}$. The singular values ($\hat{\lambda}$) and sequential sum of squares will be the same for either solution. Predicted values ($\hat{y}_{ij}$) will differ, but cultivar differences within any particular site will be the same under either solution.

Constrained LS+1 Non-COI Solution for the SREG$_2$ Model

If the $\hat{\gamma}_{jl}$ are to be forced to zero for e_1 sites in set S_1 and left unconstrained in the complementary set S_2 consisting of $e_2 = e - e_1$ sites, the residual sum of squares is

$$\sum_i \sum_{j \in S_1} (\bar{y}_{ij.} - \hat{\mu}_j)^2 + \sum_i \sum_{j \in S_2} (\bar{y}_{ij.} - \hat{\mu}_j - \hat{\lambda}_1 \hat{\alpha}_{il} \hat{\gamma}_{jl})^2$$

(Crossa and Cornelius, 1997). Both terms in this expression are minimized by putting $\hat{\mu}_j = \bar{y}_{j.}$ with $\hat{\lambda}_1$, $\hat{\alpha}_{il}$, and $\hat{\gamma}_{jl}$ for sites in set S_2 obtained as the first component of the SVD of the $g \times e_2$ matrix of deviations of cell means from site means, $\bar{y}_{ij.} - \bar{y}_{j.}$, in set S_2.

In practice, one makes a first choice of a site, which we will denote as Site k, to have its $\hat{\gamma}_{jl}$ forced to zero, i.e., as a first choice for a site to be put into set S_1. This choice will be made as described for choosing Site k for the constrained SVD non-COI solution. If the $\hat{\gamma}_{jl}$ for the $e - 1$ sites remaining in set S_2 includes values differing in sign, choose a second site to have its $\hat{\gamma}_{jl}$ forced to zero. Continue this process until the SVD of set S_2 gives the remaining nonzero $\hat{\gamma}_{jl}$ all having the same sign.

The fitted non-COI SREG$_{LS+1}$ model is obtained by means of the non-COI LS SREG$_1$ solution as the first multiplicative term and then extracting a second multiplicative term as the first component of the SVD of the matrix of deviations of the cell means from the non-COI LS SREG$_1$ solution. Vectors of $\hat{\gamma}_{j1}$ and $\hat{\gamma}_{j2}$ values appear to be orthogonal to one another, but this does not hold for vectors of $\hat{\alpha}_{i1}$ and $\hat{\alpha}_{i2}$ values.

SREG Model Using Mandel's Solution

The biplot obtained from the SREG model with Mandel's solution has been recently suggested by Yan et al. (2001) and consists in plotting, as primary effect, Mandel's solution for site regression and the first principal component extracted from the regression deviations as the secondary effect (SREG$_{M+1}$). The SREG$_{M+1}$ model is $\bar{y}_{ij} = \mu_j + b_i g_i + \lambda_1 \alpha_{i1} \delta_{j1} + \bar{\varepsilon}_{ij}$ where b_i is the regression coefficient of the jth site on the cultivar main effects (g_i) and the other terms defined as in previous cases. This equation is Mandel's sites regression ($\bar{y}_{ij.} = \mu_j + b_i g_i + \bar{\varepsilon}_{ij.}$), plus one additional multiplicative term ($\lambda_1 \alpha_{i1} \delta_{j1}$) estimated by subjecting the matrix of deviations from the Mandel's regression model ($\bar{y}_{ij.} - \mu_j - b_i g_i$) to SVD.

Biplot

Biplots obtained from linear-bilinear models, such as SHMM and SREG, are constructed from the SVD of the two-way table of deviations of empirical cell means from least squares estimates of the additive components. On a two-dimensional Cartesian coordinate system, *markers* for cultivars are plotted with primary effect (score in first multiplicative term) and secondary effect (score in second multiplicative term) as coordinates. A set of markers for sites is plotted on the same figure, also with primary and secondary effects as coordinates.

A full description of the interpretation of the biplots of multiplicative models is given in Gower and Hand (1996). Briefly, the cultivar and site scores are represented as vectors in a two-dimensional space, so it is useful to interpret biplots in terms of directions of the vectors and their projections. Cultivar and site vectors are defined as vectors from the origin (0,0) to the end points determined by their markers (scores). An angle $\theta < 90°$ or $\theta > 270°$ between a cultivar vector and a site vector indicates that the cultivar had a positive response at that site. A negative cultivar response is indicated if $90° < \theta < 270°$. Note that in the SREG model, the interpretation of the biplot is with respect to the variation for which main effects of cultivars (G) and the GEI (G+GEI) account, whereas in the SHMM biplot, the interpretation is on the deviations from the shift parameter. Performance of a cultivar in a site can be approximated by the orthogonal projection of the cultivar vector onto the line determined by the direction of the site vector; that is, if we consider the line containing the site vector, the cultivar's response at that site is approximated by the length of the segment of that line extending from the origin to the point where that line can be perpendicularly intersected by a line drawn from the cultivar marker.

The cosine of the angle between two site (or cultivar) vectors approximates the phenotypic correlation of yield performance of the two sites (or cultivars). An angle of zero indicates a correlation of +1; an angle of 90° (or −90°), a correlation of 0; and an angle of 180°, a correlation of −1. Furthermore, the cultivar scores for the first multiplicative component of the SREG model will usually be closely associated with the cultivar main effects.

The biplot methodology of Yan et al. (2000) forms a polygon by joining the most extreme cultivars of the biplot with line segments, one for each side of the polygon drawn from the origin to perpendicularly intersect that side of the polygon. These perpendiculars are further extended sufficiently far to subdivide the biplot into sectors so that each site marker and each cultivar marker is contained within one (and only one) sector. When a polygon cannot be formed because primary effects of cultivars, as well as primary effects of sites, are all of the same sign, but the signs for cultivars are opposite to those for sites, one can still draw straight lines joining the most extreme cultivars to form a polygon, as well as lines that pass through the origin and are perpendicular to the sides of the polygon. In many cases, perpendicular lines from the center of the biplot are drawn, but their intersection falls on the extension of the side of the polygon beyond the corner (vertex) where the side ends.

Rescaling the Singular Vectors

For the graphical display of the biplots, it is advisable to absorb the singular values of the first and second multiplicative

components, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, into the singular vectors of sites ($\hat{\gamma}_1$ and $\hat{\gamma}_2$) and cultivars ($\hat{\alpha}_1$ and $\hat{\alpha}_2$) in such a way that the products of rescaled primary and secondary effects are equal to the contributions $\hat{\lambda}_1\hat{\alpha}_{i1}\hat{\gamma}_{j1}$ and $\hat{\lambda}_2\hat{\alpha}_{i2}\hat{\gamma}_{j2}$ of the first and second multiplicative components, respectively, to the predicted values of the attribute.

Let the rescaled values of the singular vectors of cultivars and sites be $\hat{\alpha}_1^* = \hat{\lambda}_1^A \hat{\alpha}_1$ and $\hat{\gamma}_1^* = \hat{\lambda}_1^{(1-A)} \hat{\gamma}_1$, respectively, with $0 \le A \le 1$. Select a value of A, such that it will force the range of values in the singular vector for sites to be equal to the range of values in the singular vector for cultivars, that is, $\max(\hat{\alpha}_1^*) - \min(\hat{\alpha}_1^*) = \max(\hat{\gamma}_1^*) - \min(\hat{\gamma}_1^*)$. Define C = $\max(\hat{\alpha}_1) - \min(\hat{\alpha}_1)$ and D = $\max(\hat{\gamma}_1) - \min(\hat{\gamma}_1)$ (where max and min denote the largest and smallest elements of the vector, respectively). Then choose A such that $C(\hat{\lambda}_1^A) = D(\hat{\lambda}_1^{1-A})$ or equivalently, $\hat{\lambda}_1^{2A-1} = D/C$. Solving for A gives $(2A - 1)\log(\hat{\lambda}_1) = \log(D/C)$ and then

$$A = \frac{1}{2}\left[1 + \left\{\frac{\log(D/C)}{\log(\hat{\lambda}_1)}\right\}\right]$$

and

$$(1 - A) = \frac{1}{2}\left[1 - \left\{\frac{\log(D/C)}{\log(\hat{\lambda}_1)}\right\}\right].$$

If the ranges are the same, i.e., $D = C$, then $A = (1 - A) = 0.5$. Value for A for rescaling the vectors in the second multiplicative term is computed similarly.

Defining Levels of COI

It is useful to classify subsets of sites and cultivars with different levels of COI, defined in terms of how much rank displacement has occurred in the COI. An *order h−1 adjacent COI* subset will be defined as a subset of e_1 sites and a subset of h cultivars such that the ranks of this subset of cultivars in the ranking of all cultivars in each of the sites in the subset is some permutation of the integers $r+1$, $r+2$,..., $r+h$, with these permutations not being the same at all sites in the subset, but with r being a constant integer, $0 < r \le g - h$. The *level* of the adjacent COI subset will be defined as the maximum cultivar rank change that occurs from one site to another in the subset. Our main interest will be the cases where $r = 0$ and $h = 2$ (or 3), which constitute cases where the best two (or three) cultivars are the same two (or three) cultivars in every site in the subset. In other words, we will be interested in *order 1 adjacent COI* and *order 2 adjacent COI*.

For our purposes in this paper, an adjacent subset will be considered *low level* if its order is ≤2. Thus, since the level of an adjacent COI subset cannot exceed its order, cases where $h = 2$ or 3 are necessarily low level. In the sequel, for brevity we will drop the word *adjacent*, and simply characterize such subsets as *low level COI subsets*. Note that a level 0 *adjacent* COI subset cannot exist, because there must be at least two cultivars with rank changes if more than one permutation of the subset of ranks exists. Thus, only a non-COI subset can be at level 0 with respect to cultivar rank changes. In our usage, for a subset of cultivars to be *adjacent*, the members of the subset must not only be "adjacent" in every site, but they also must be consistently *adjacent*, i.e., r must be constant. When r is not constant but differs from site to site, the subset of cultivars is inconsistently *adjacent* in the subset of sites. (An extreme case would be when three cultivars are the three best in some sites and the three worst in other sites.)

Software

Unconstrained and constrained SVD non-COI $SREG_2$ and $SHMM_2$ solutions can be computed by the FORTRAN program EIGAOV that can be run on a personal computer. The constrained LS+1 solution for SREG was obtained by importing the constrained non-COI LS $SREG_1$ *solution from* EIGAOV into SAS/IML (SAS Institute, Inc., 1989) to complete the computation. Information about the use of the EIGAOV programs can be obtained from the second author.

RESULTS AND DISCUSSION

Trial 1

Standard errors of the cultivar means ranged from 142 kg ha^{-1} (Site 10) to 722 kg ha^{-1} (Site 19) (Table 1). The Bartlett's test rejected the hypothesis of homogeneous within site error variance, and the Shapiro-Wilk test for non-normality of residuals at each site indicated that the normality assumption is acceptable for all sites. Cultivars G4, G5, and G6 had low level COI (ranked within the best three cultivars) in Sites 4, 5, 13, 14, 15, and 16. Similarly, there was low-level COI between Cultivars G3 and G9 in Sites 2, 4, 7, 13, 14, 15, 16, and 17 (Cultivars G3 and G9 ranked between 4th and 6th). Cultivars G1, G2, G7, and G8 ranked among the worst five cultivars in Sites 4, 5, 9, 11, 12, 13, 14, 15, 16, and 20. In Site 1, the best cultivar was G8, whereas in Site 8, the best cultivar was G1, so Cultivars G8 and G1 in Sites 1 and 8 showed a high level COI.

Rankings of the cultivar predicted values in each site for $SHMM_2$ and $SREG_2$ models based on scaled and unscaled data, and for the different non-COI constrained solutions are presented in Tables A1 and A2 (Appendix 2).

Unconstrained and Constrained $SREG_2$ Solutions and Their Biplots

For the unscaled data and unconstrained model, $SREG_2$(U/U), the F_R test of Cornelius et al. (1992), which assesses the significance of the residual variation after fitting the first $k - 1$ multiplicative components, found no significant residual ($P \le 0.05$) after fitting the second multiplicative component, whereas the F_{GH1} test (Cornelius et al., 1996) used for judging the significance of sequentially fitted multiplicative terms found three significant terms ($P \le 0.05$). For $SREG_2(U/C_{SVD})$, three and four significant terms were found significant ($P \le 0.05$) by the F_{GH1} and F_R tests, respectively. For the scaled data and unconstrained model, $SREG_2$(S/U), three and four terms were significant ($P \le 0.05$) by the F_{GH1} and F_R tests, respectively, and for $SREG_2(S/C_{SVD})$, both tests found four significant terms ($P \le 0.05$).

The biplot of the $SREG_2$ model, using unscaled data, $SREG_2$(U/U) (Fig. 1A) shows that cultivars, based on the sign of their primary effects ($\hat{\alpha}_{i1}$), are divided into two groups, {G1, G2, G3, G7, G8} vs. {G4, G5, G6, G9}. Sites, based on the sign of their primary effects ($\hat{\gamma}_{j1}$), are divided into two groups {1, 3, 8, 10} vs. {2, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}. Cultivars G1, G2, G3, G7, and G8 have a positive response in terms of their primary effects and GEI at Site 8 (their orthogonal projections onto the line containing the site vector are in the same direction as the site vector) as opposed to

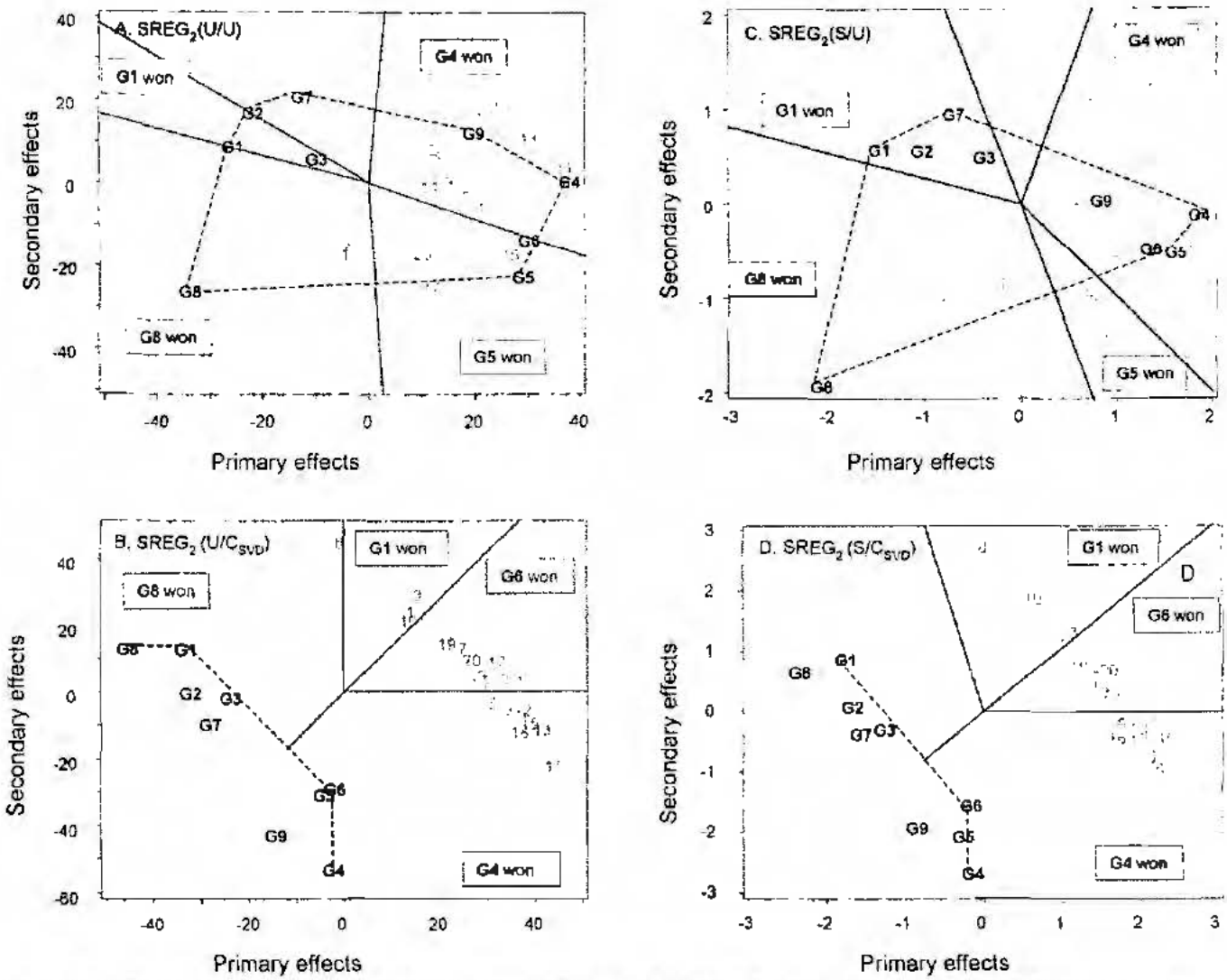

Fig. 1. Trial 1 biplots of: (A) $SREG_2$(U/U) = sites regression model on unscaled data and unconstrained solution; (B) $SREG_2$(U/C_{SVD}) = sites regression model on unscaled data and constrained SVD solution; (C) $SREG_2$(S/U) = sites regression model on scaled data and unconstrained solution; (D) $SREG_2$(S/C_{SVD}) = sites regression model on scaled data and constrained SVD solution.

the projection of Cultivars G4, G5, G6, and G9 that are located in the opposite side.

Since the correlation between the primary effects of cultivars and their main effects is high (0.97), the biplot can be used for cultivar evaluation with respect to performance ability ($\hat{\alpha}_{i1}$) and stability ($\hat{\alpha}_{i2}$) and site evaluation with respect to discrimination ($\hat{\gamma}_{j1}$) and representativeness ($\hat{\gamma}_{j2}$) (Yan et al., 2001). In Fig. 1A, genotype G4 has the largest primary effect (high mean yield) and near-zero secondary effect (stable across most sites), whereas Sites 1 and 3 do not discriminate cultivars well (relatively small $\hat{\gamma}_{j1}$), and relatively large $\hat{\gamma}_{j2}$ values predict large GEI, that is, inconsistency of cultivar responses in Sites 1 and 3 as compared with their responses in other sites.

The polygon has seven vertices located at markers for Cultivars G1, G2, G7, G9, G4, G5, and G8. In the upper right sector of Fig. 1A, Cultivar G4 had the best $SREG_2$ predicted values at Sites 2, 4 through 6, 9, 11 through 14, 16, and 18 (Table 3), but it ranked among the worst three cultivars (7th, 8th, and 9th ranks) in sites located in the opposite sectors containing Sites 1, 3, 8, and 10 (Table 3). On the contrary, Cultivar G8 is the winner in Sites 1, 3, and 10 and the loser in sites located in the opposite sector, Sites 2, 4 through 6, 9, 11 through 14, 16, and 18. Thus, Cultivars G4 vs. G8 and Sites 1, 3, 8 and 10 (with negative primary effects) vs. Sites 2, 4 through 6, 9, 11 through 14, 16, and 18 (with positive primary effects and located in opposite sector of the biplot) had a clear COI pattern. Cultivars G1 and G5 are the winner (1st rank) and loser (9th rank) in Site 8, respectively (Table 3). Similar patterns of COI can be observed for the cultivar subset {G1 and G2} vs. G5 at Sites 7, 15, 17, 19, and 20 (with positive primary effects) as compared to Site 8 (with negative primary effects). The Site 8 marker was located far away from the other site markers so that it can be considered very different from the other sites.

The vertex cultivars located in opposite sectors of the biplot, along with the sites included in those sectors,

Table 3. Crossover interaction subsets of cultivars and sites represented by the winner (1st rank) and loser (7th, 8th, and 9th ranks) cultivars in specific sites based on predicted values for each type of model-data-constraint combination for Trial 1 data.

Cultivar	Sites where cultivar is: Winner	Loser
	$SREG_2(U/U)$†	
G1	8	2,4–7,9,11–20
G4	2,4–6,9,11–14,16,18	1,3,8,10
G5	7,15,17,19,20	8
G8	1,3,10	2,4–6,9,11–14,16,18
	$SREG_2(U/C_{SVD})$	
G1	1,3,10	2,4–6,9,11–18
G4	4,5,11–13,15,18	1,3,8,10
G6	2,6,7,9,14,16,17,19,20	–
G8	8	2,4–7,9,11–20
	$SREG_2(S/U)$	
G1	8	2,4–7,9,11–20
G4	2,4–7,9,11–16,18,19	1,3,8,10
G5	17,20	3,8,10
G8	1,3,10	2,4–6,9,11–16,18
	$SREG_2(S/C_{SVD})$	
G1	1,3,8,10	2,4–6,11–13,15–18,20
G4	2,3,4,11–13,15,16,18	1,3,8,10
G6	6,7,9,14,17,19,20	–
	$SREG_2(U/C_{LS+1})$	
G1	3,8,10	1,2,4,5,7,9,11,12,14–20
G4	1,2,4–7,9,11–20	3,8,10
	$SREG_2(S/C_{LS+1})$	
G1	1,3,8,10	2,4–5,7,9,11,12,14–20
G4	2,4–7,9,11–20	1,3,8,10
	$SREG_{M+1}(U)$	
G4	2,5,11,12,14,18	1,3,8,10
G5	3,7,17,20	–
G6	4,6,9,13,15,16,19	1,8,10
G8	1,8,10	2,4–7,9,11–20

Cultivar	Sites where cultivar is: Winner	Loser
	$SREG_{M+1}(S)$	
G1	3,8	1,4,5,7,11,14,15,17,19,20
G4	17,19	3,8,10
G6	2,4–7,9,11–16,18,20	1,8,10
G8	1,10	2–4,6,7,9,11–16,18
	$SHMM_2(U/U)$	
G1	3,8	2,4,5,7,9,11,13,15–19
G4	2,4,5,7,9,11,13,15–18,20	1,3,8,10
G5	6,12,14	1,10
G8	1,10	2,4–6,9,11–18,20
G9	19	3,8
	$SHMM_2(U/C_{SVD})$	
G1	1,3,8,10	2,4,5,7,9,11,13,15–20
G4	2,4,5,7,9,11,13,15–19	1,3,8,10
G5	6,12,14,20	–
	$SHMM_2(S/U)$	
G1	1,3,8,10	2,4,5,7,9,11–13,15–20
G4	2,4,5,7,9,11–13,15–20	1,3,8,10
G5	6,14	–
	$SHMM_2(U/C_{LS+1})$	
G1	1,3,8,10	2,4,5,7,9,11,13,15–19
G4	2,4,5,7,9,11,13,15–19	1,3,8,10
G5	6,12,14,20	–

† $SREG_2(U/U)$ = sites regression model on unscaled data and unconstrained solution;
$SREG_2(U/C_{SVD})$ = sites regression model on unscaled data and constrained SVD solution;
$SREG_2(S/U)$ = sites regression model on scaled data and unconstrained solution;
$SREG_2(S/C_{SVD})$ = sites regression model on scaled data and constrained SVD solution;
$SREG_2(U/C_{LS+1})$ = sites regression model on unscaled data and constrained LS + 1 solution;
$SREG_2(S/C_{LS+1})$ = sites regression model on scaled data and constrained LS + 1 solution;
$SREG_{M+1}(U)$ = Mandel's sites regression model on unscaled data and unconstrained solution;
$SREG_{M+1}(S)$ = Mandel's sites regression model on scaled data and unconstrained solution;
$SHMM_2(U/U)$ = shifted multiplicative model on unscaled data and unconstrained solution;
$SHMM_2(U/C_{SVD})$ = shifted multiplicative model on unscaled data and constrained SVD solution;
$SHMM_2(S/U)$ = shifted multiplicative model on scaled data and unconstrained solution;
$SHMM_2(U/C_{LS+1})$ = shifted multiplicative model on unscaled data and constrained LS + 1 solution.

formed COI subsets of cultivars and sites. Detecting low level COI and non-COI groups, however, does not require development of the polygon, but rather identification of subsets of cultivar and site markers with the same directions. For example, in Fig 1A, Cultivars G4, G5, and G6 are the best three performers in Sites 4, 5, 7, 9, 11, 14 through 17, 19, and 20 (essentially the lower right quadrant) (Table 4), closely followed by Cultivar G9; these cultivars had projections onto the positive directions of site vectors for those sites, but they were the worst three cultivars in Site 8; they project onto the negative direction (opposite quadrant) of Site 8. Thus, Cultivars G4, G5, and G6 and Sites 4, 5, 7, 9, 11, 14 through 17, 19, and 20 formed a clear low level COI subset. Other subsets of sites with non-COI for all cultivars are Sites 2, 6, 12, 13, and 18 (upper right quadrant of Fig. 1A) and Sites 1 and 3 (lower left quadrant of Fig. 1A) (Table A1, Appendix 2).

In general, the ranking of the $SREG_2(U/U)$-predicted values of Table A1 (Appendix 2) approximates the ranking of the observed values of Table 1, although some distortions are noteworthy, e.g., the observed value of Cultivar G4 that ranked 3rd at Site 1, whereas its $SREG_2$-predicted value ranked 7th and Cultivar G5 that ranked 8th at Site 1, but its $SREG_2$-predicted value ranked 2nd. This result would suggest that, because Site 1 has small cultivar differences, data from it will fit practically any multiplicative model used. A similar statement can be made concerning Site 7 vis-a-vis Cultivars G4 and G7.

The biplot of the constrained $SREG_2$ model using unscaled data and SVD non-COI constrained solution, $SREG_2(U/C_{SVD})$ (Fig. 1B), showed Sites 1, 3, and 10 with $\hat{\gamma}_{j1} > 0$ and Site 8 with $\hat{\gamma}_{j1} = 0$. Similar to the $SREG_2(U/U)$ model, two groups of cultivars are formed {G1, G2, G3, G7, G8} and {G4, G5, G6, G9}. Constraint of the first term of $SREG_2$ gave all primary effects for cultivars with negative values and all primary effects for sites with non-negative values (zero for Site 8). The lower dispersion of the points in this biplot reflects the lower variability explained by the constrained solution as compared with that obtained by the unconstrained

Table 4. Low level crossover interaction (COI) and non-COI subsets of cultivars and sites based on predicted values for each type of model-data-constraint combination for Trial 1 data.

Cultivar	Sites where the cultivars have low level COI or non-COI
$SREG_2(U/U)$†	
G4,G5,G6	4,5,7,9,11,14–17,19,20
G4,G6,G9	2,6,12,13,18
G5,G6,G8	1,3
$SREG_2(U/C_{SVD})$	
G4,G5,G6	2,4–7,9,11–20
G1,G3	1,3,10
$SREG_2(S/U)$	
G4,G5,G6	2,4,5,7,9,11–20
G1,G2,G8	3,10
$SREG_2(S/C_{SVD})$	
G4,G5,G6	2,4–7,9,11–20
G1,G3,G6	1,3
G1,G2,G8	8,10
$SREG_2(U/C_{LS+1})$	
G4,G5,G6	1,2,4–7,9,11–20
G1,G2,G3	3,8,10
$SREG_2(S/C_{LS+1})$	
G4,G5,G6	2,4–7,9,11–20
G1,G2,G3	1,3,8,10
$SREG_{M+1}(U)$	
G4,G5,G6	4,6,7,9,13–17,19,20
G1,G8	1,8,10
G4,G6,G9	2,5,11,12,18
$SREG_{M+1}(S)$	
G4,G5,G6	2,4–7,9,11–20
$SHMM_2(U/U)$	
G4,G5,G6	2,4–6,9,11–18,20
G1,G2	3,8
$SHMM_2(U/C_{SVD})$	
G4,G5,G6	2,4–7,9,11–18,20
G1,G2,G3	1,3,10
$SHMM_2(S/U)$	
G4,G5,G6	2,4–7,9,11–20
G1,G2	1,3,10
$SHMM_2(U/C_{LS+1})$	
G4,G5,G6	2,4–7,9,11–20
G1,G2,G3	1,3
G1,G2	8,10

† $SREG_2(U/U)$ = sites regression model on unscaled data and unconstrained solution;
$SREG_2(U/C_{SVD})$ = sites regression model on unscaled data und constrained SVD solution;
$SREG_2(S/U)$ = sites regression model on scaled data and unconstrained solution;
$SREG_2(S/C_{SVD})$ = sites regression model on scaled data and constrained SVD solution;
$SREG_2(U/C_{LS+1})$ = sites regression model on unscaled data and constrained LS + 1 solution;
$SREG_2(S/C_{LS+1})$ = sites regression model on scaled data and constrained LS + 1 solution;
$SREG_{M+1}(U)$ = Mandel's sites regression model on unscaled data and unconstrained solution;
$SREG_{M+1}(S)$ = Mandel's sites regression model on scaled data and unconstrained solution;
$SHMM_2(U/U)$ = shifted multiplicative model on unscaled data and unconstrained solution;
$SHMM_2(U/C_{SVD})$ = shifted multiplicative model on unscaled data and constrained SVD solution;
$SHMM_2(S/U)$ = shifted multiplicative model on scaled data and unconstrained solution;
$SHMM_2(U/C_{LS+1})$ = shifted multiplicative model on unscaled data and constrained LS + 1 solution.

solution. Although a polygon that contains the plot origin (0, 0) cannot be drawn, the properties of the biplot remained the same as those given for the biplot obtained with the unconstrained solution. Figure 1B predicts COI for Cultivars G1 and G4 at Sites 1, 3, and 10 as compared to Sites 4, 5, 11 through 13, 15, and 18 (Table 3). Similarly, COI is found between Cultivars G6 and G8 in Sites 2, 6, 7, 9, 14, and 16 through 20 as compared with Site 8. The Site 8 marker is far away from the others in the biplot; the constrained solution sets its primary effects equal to zero, and 50% of the variation explained by the second multiplicative component is due to cultivar differences within Site 8. The line perpendicular to the segment joining Cultivars G1 and G6 separates the two non-COI groups of sites and cultivars. One low order COI subset comprises Cultivars G4, G5, and G6 and all sites except Sites 1, 3, 8, and 10 (note that Cultivar G9 ranked fourth in most of these sites) (Table 4). A non-COI group includes Cultivars G1 and G3 and Sites 1, 3, and 10 (Table 4).

For the unconstrained $SREG_2$ model using scaled data, $SREG_2(S/U)$, the biplot (Fig. 1C) gave results similar to those found for $SREG_2(U/U)$. In the right sector, Cultivar G4 had the best $SREG_2$ predicted values at Sites 2, 4 through 7, 9, 11 through 16, 18, and 19 (Table 3), but it ranked among the worst three cultivars (7th, 8th, and 9th ranks) in sites located in the opposite sectors, Sites 1, 3, 8, and 10 (Table 3). On the contrary, Cultivar G8 is the winner in Sites 1, 3, and 10 and the loser in sites located in the opposite sector, Sites 2, 4 through 6, 9, 11 through 16, and 18. Thus, Cultivars G4 and G8 show COI at Sites 2, 4 through 6, 9, 11 through 16, and 18 (with positive primary effects) as compared to Sites 1, 3, and 10 (with negative primary effects). On the other hand, Cultivars G4, G5, and G6 and Sites 2, 4, 5, 7, 9, and 11 through 20 represent a low level COI group (Table 4); these cultivars with Sites 3, 8, and 10 formed a non-COI group but in the negative direction (poor yield response). Also, Cultivars G1, G2, and G8 and Sites 3 and 10 formed a non-COI group.

The biplot of the $SREG_2(S/C_{SVD})$ (Fig. 1D) is similar to the $SREG_2(U/C_{SVD})$ biplot. Cultivars G1 vs. G4 and Sites 2, 4, 5, 11 through 13, 15, 16, and 18 vs. Sites 1, 3, 8, and 10 formed a COI group (Table 3). Cultivars G4, G5, and G6 were the three best ranking cultivars in all sites except Sites 1, 3, 8, and 10, and thus formed a low level COI subset. Cultivars G1, G3, and G6 in Sites 1 and 3 and also Cultivars G1 and G8 in Sites 8 and 10 (Table 4) formed a non-COI group. Site 8 is very distinct from the others and explained 51% of the second term variability (data not shown). Biplot of the constrained non-COI solution showed less dispersion of points but interpretation similar to the biplots obtained from unconstrained solutions.

Constrained LS+1 $SREG_2$ Solution and its Biplot. In the biplot of the LS+1 constrained $SREG_2$ model using unscaled data, $SREG_2(U/C_{LS+1})$ (Fig. 2A) produced a polygon that is a triangle in which Cultivars G1 and G4 have a COI in Sites 3, 8, and 10 as compared to the rest of the sites (Table 3). Most of the variation described by the second component (80%) is due to cultivar differences within Site 8. Sites 1, 3, and 10 are located toward the center of the biplot and thus cultivar differences at those sites are small. Cultivars, based on the sign of $\hat{\alpha}_{i1}$,

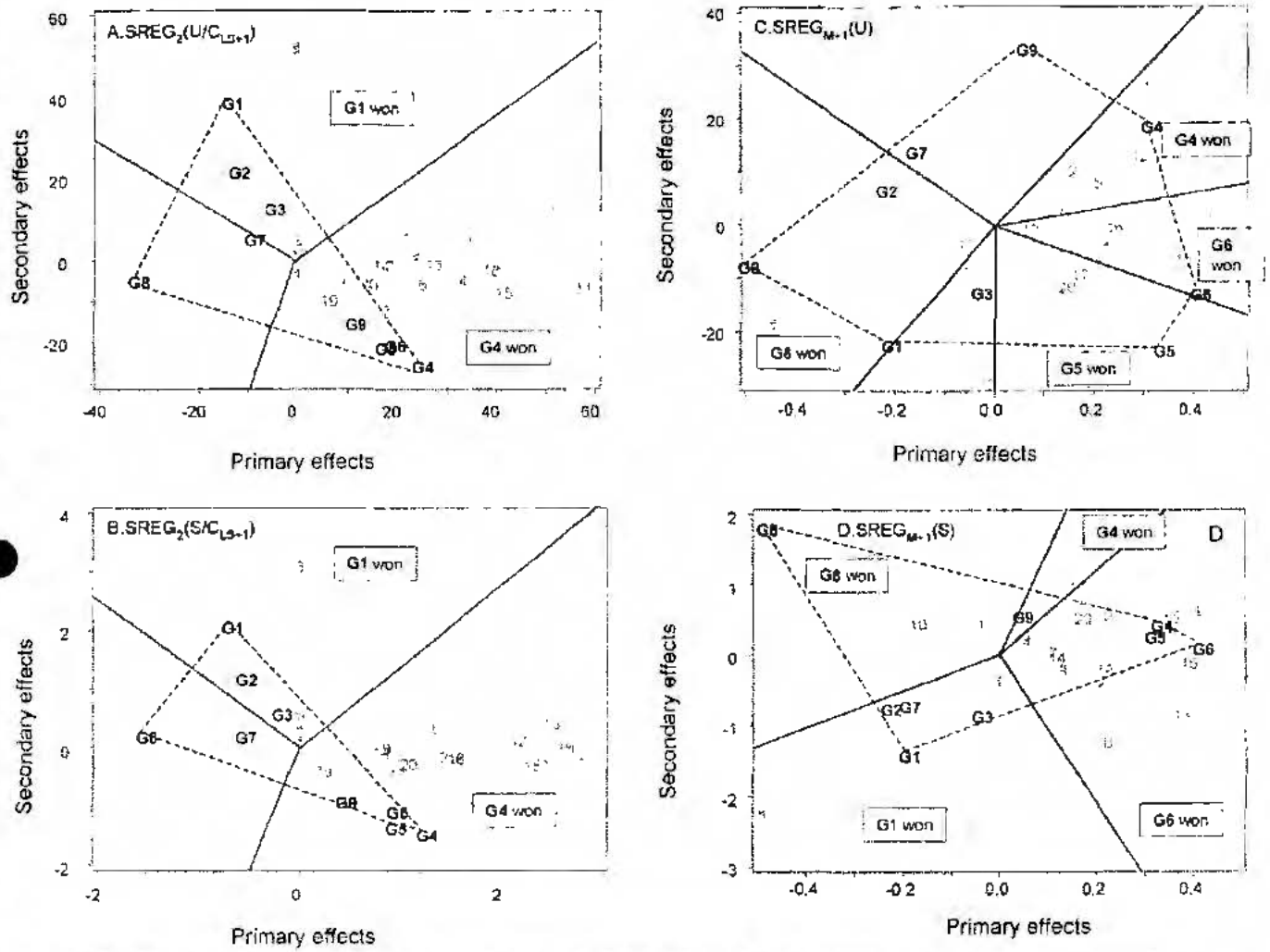

Fig. 2. Trial 1 biplots of: (A) $SREG_2(U/C_{LS+1})$ = sites regression model on unscaled data and a constrained LS solution; (B) $SREG_2(S/C_{LS+1})$ = sites regression model on scaled data and constrained LS solution; (C) $SREG_{M+1}$ (U) = Mandel's sites regression model on unscaled data and unconstrained solution; (D) $SREG_{M+1}(S)$ = Mandel's sites regression model on scaled data and unconstrained solution.

are divided into two groups, {G1, G2, G3, G7, G8} vs. {G4, G5, G6, G9}. Sites 1, 3, 8, and 10 have $\hat{\gamma}_{ji} = 0$. Cultivars G4, G5, G6, and G9 in all sites, except Sites 3, 8, and 10, formed a clear low level COI subset and Cultivars G1, G2, G3, G7, and G8 in Sites 3, 8, and 10 (Table 4) formed a non-COI subset. Very similar COI subsets of cultivars and sites are found for the biplot on the scaled data, $SREG_2(S/C_{LS+1})$ (Fig. 2B), except that now Cultivar 8 ranked fourth in Sites 1, 3, 8, and 10 and that for Site 1 the best three cultivars were G1, G2, and G3.

In this constrained solution, the first and second singular vectors for sites are orthogonal to each other, but the singular vectors for cultivars seem to have a negative linear association, reflecting the strong COI involving Site 8 and, to a lesser extent, 3 and 10 versus the rest with respect to the complete predictive rank reversal of cultivar sets {G1, G2, G3, and G7} vs. {G4, G5, G6, and G9}.

SREG Model Using Mandel's Solution and its Biplot. Recently, Yan et al. (2001) showed that the biplots from the SREG model using the Mandel solution ($SREG_{M+1}$) and the standard $SREG_2$ model gave similar winning cultivars as well as GEI interaction patterns. The advantage of the $SREG_{M+1}$ biplot is that the first component indicates mean yield and the second component stability; for the $SREG_2$ model, this is so only if the first bilinear component is highly correlated with the cultivar main effects.

The biplot of the $SREG_{M+1}$ model using unscaled data, $SREG_{M+1}(U)$ (Fig. 2C) showed the same split of cultivars and sites that was previously found, that is, {G1, G2, G3, G7, G8} vs. {G4, G5, G6, G9} and sites {1, 3, 8, 10} vs. the rest. A polygon with six vertices G1, G5, G6, G4, G9, and G8 (counter-clockwise around the polygon) is formed with Cultivars G4 and G8 at opposite sectors having COI in Sites 1, 8, and 10 as compared with Sites 2, 5, 11, 12, 14, and 18 (Table 3). Cultivars G4, G5 and G6 had the best three predicted values in Sites 4, 6, 7, 9, 13 through 17, 19, and 20 (located toward the lower right quadrant of the biplot) and thus formed a low level COI group (Table 4). Cultivars G1 and G8 are the best two in Sites 1, 8, and 10 and formed a non-COI group (Table 4 and Appendix 2 Table A1). Sites

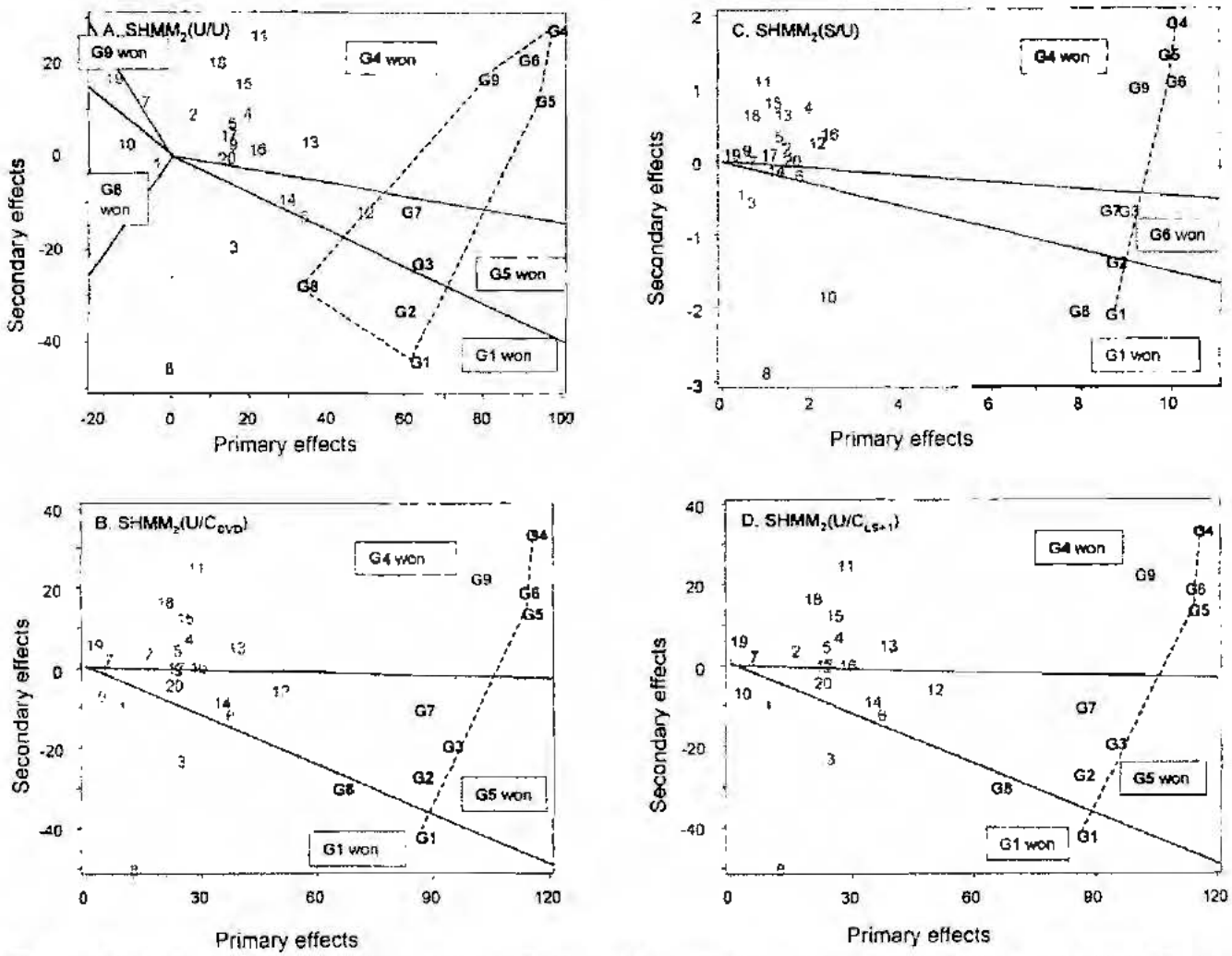

Fig. 3. Trial 1 biplots of: (A) SHMM$_2$(U/U) = shifted multiplicative model on unscaled data and unconstrained solution; (B) SHMM$_2$U/C$_{SVD}$) = shifted multiplicative model on unscaled data and a constrained SVD solution; (C) SHMM$_2$(S/U) = shifted multiplicative model on scaled data and unconstrained solution; (D) SHMM$_2$(U/C$_{LS+1}$) = shifted multiplicative model on unscaled data and a constrained LS solution.

2, 5, 11, 12, and 18 had very similar cultivar ranking and gave Cultivars G4, G6 and G9 as the best three performers and thus formed a low level COI group.

The biplot of the SREG$_{M+1}$ model using scaled data SREG$_{M+1}$(S) (Fig. 2D) showed a COI pattern between Cultivars G8 and G6 in Sites 1 and 10 as compared to Sites 2, 4, 6, 7, 9, 11 through 16, 18 (Table 3). Similar to the SREG$_{M+1}$(U) case, Cultivars G4, G5, and G6 had the best three predicted values in Sites 2, 4 through 7, 9, 11 through 20 and, thus, formed a clear low level COI group (Table 4).

Unconstrained and Constrained SHMM$_2$ and Their Biplots

For the SHMM$_2$(U/U) and SHMM$_2$(U/C$_{SVD}$), the F_R and F_{GH1} tests found that the first three multiplicative components were significant ($P \leq 0.05$). For SHMM$_2$-(S/U), the F_R and F_{GH1} tests found that the first three and four multiplicative components were significant ($P \leq 0.05$), respectively. Since all primary effects of sites for SHMM$_2$(S/U) model were of the same sign, biplots with non-COI constrained SHMM$_2$ solutions for this model were not required. The biplot of the SHMM$_2$-(U/U) (Fig. 3A) model had a subset of sites {1, 7, 8, 10, and 19} with negative values of site primary effects, while the rest of the sites had positive values. All cultivars had positive and high values for their primary effects. The secondary effects of cultivars separated them into two groups {G1, G2, G3, G7, and G8} vs. {G4, G5, G6, and G9}. This subdivision of cultivars was also obtained for SHMM$_2$(U/C$_{SVD}$) (Fig. 3B), SHMM$_2$(S/U) (Fig. 3C), and SHMM$_2$(U/C$_{LS+1}$) (Fig. 3D).

All the biplots of the SHMM$_2$ model indicated a COI pattern between most distant cultivars in the biplot, G1 and G4 in Sites 1, 3, 8, and 10 as compared to Sites 2, 4, 5, 7, 9, 11, 13, 15 through 19 (Table 3). Low level COI and non-COI patterns between cultivars and sites that are located toward the upper region of all SHMM$_2$ biplots, as compared to those sites and cultivars located in the lower region, can be identified. For example, in all the biplots, Cultivars G4, G5, and G6 ranked within the best three predicted performers in Sites 2, 4 through

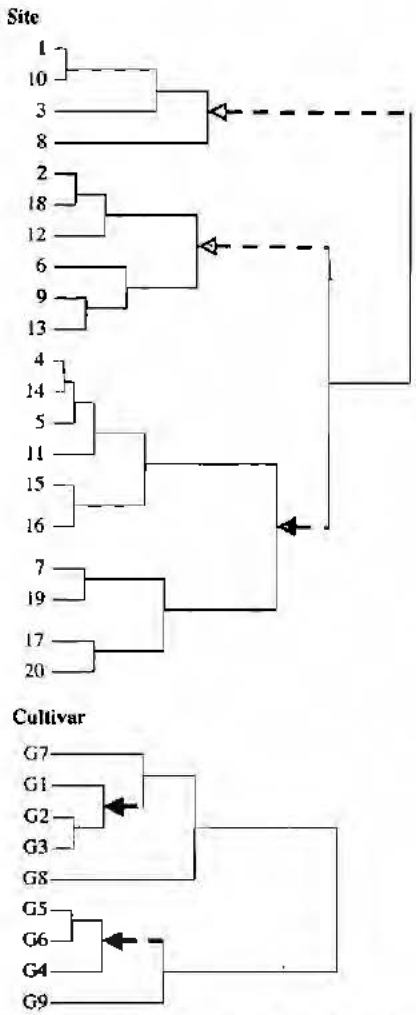

Fig. 4. Dendrograms resulting from clustering 20 sites and nine cultivars of Trial 1 using empirical cell means as input data.

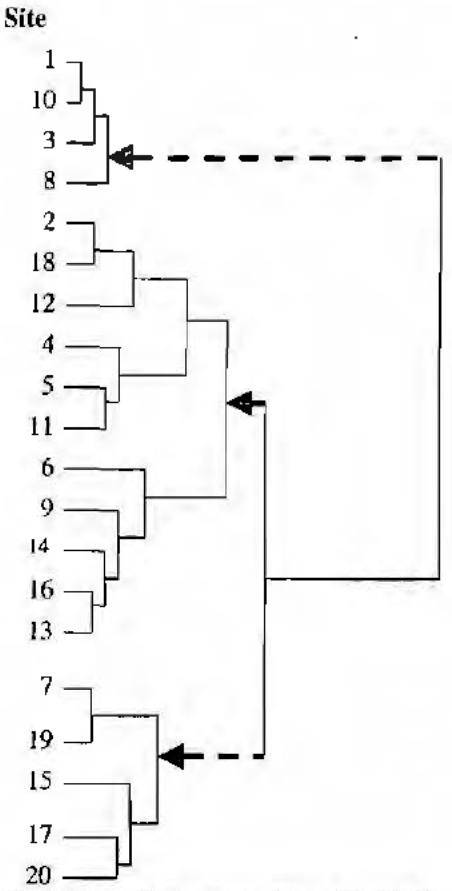

Fig. 5. Dendrogram resulting from clustering 20 sites of Trial 1 using SREG shrinkage estimates of cell means as input data.

6, 9, 11 through 18, and 20 (in most of these sites Cultivar G9 ranked fourth) and thus formed a low-level COI group (Table 4 and Appendix 2 Table A2). Note that this pattern is also found in Site 7 if the method is $SHMM_2(U/C_{SVD})$, $SHMM_2(S/U)$ or $SHMM_2(U/C_{LS+1})$, and in Site 19 if the method is $SHMM_2(S/U)$.

Cultivar G1 is chosen by $SHMM_2(U/C_{SVD})$, $SHMM_2(S/U)$, and $SHMM_2(U/C_{LS+1})$ as best in Sites 1, 3, 8, and 10 and Cultivars G1, G2, and G3 are found as a low level COI group in Sites 1, 3, and 10 by $SHMM_2(U/C_{SVD})$, in Sites 3 and 10 by $SHMM_2(S/U)$, and in Sites 1 and 3 by $SHMM_2(U/C_{LS+1})$. Also, all the $SHMM_2$ biplots indicated that Site 8 is very different from the others. It is apparent that the constrained $SHMM_2$ solutions do not affect the interpretability of the biplots for finding COI and non-COI groups of cultivars and sites.

Clustering of Sites or Cultivars into Groups with Non-COI

It is useful to investigate the clustering of sites (or cultivars) into non-COI subsets. This was done by means of the $SREG_t$ model and the clustering strategy proposed by Crossa et al. (1993) for grouping sites, and the fusion method of Crossa and Cornelius (1993) for clustering cultivars.

Recently, Cornelius et al. (1996) and Cornelius and Crossa (1999), in a cross-validation study involving five multienvironment cultivar trials, found that shrinkage estimates of multiplicative models were usually more accurate for predicting the response of cultivars within sites that were best truncated multiplicative models fitted by least squares, best linear unbiased predictors (BLUPs) based on a two-way random effects model with interaction, and the empirical cell means. For Trial 1 (Trial 3 in Cornelius and Crossa, 1999), the shrinkage estimates of multiplicative models were better predictors than BLUPs and empirical cell means. Consequently, clustering of sites (or cultivars) in Trial 1 into non-COI groups was also computed by means of distance between sites computed with the empirical cell (cultivar × site) means replaced by SREG shrinkage estimates as input data.

Dendrograms and final groups of sites and cultivars based on empirical cell means are shown in Fig. 4 and dendrogram and final groups of sites based on SREG shrinkage estimates of cell means are shown in Fig. 5. In both cases, sites are grouped into two major clusters {1, 3, 8, 10} vs. {2, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}; cultivars are split into two main subsets {G1, G2, G3, G7, G8} vs. {G4, G5, G6, G9}. This separation of the sites and cultivars into two main groups was consistently found in all model-data-constraint combi-

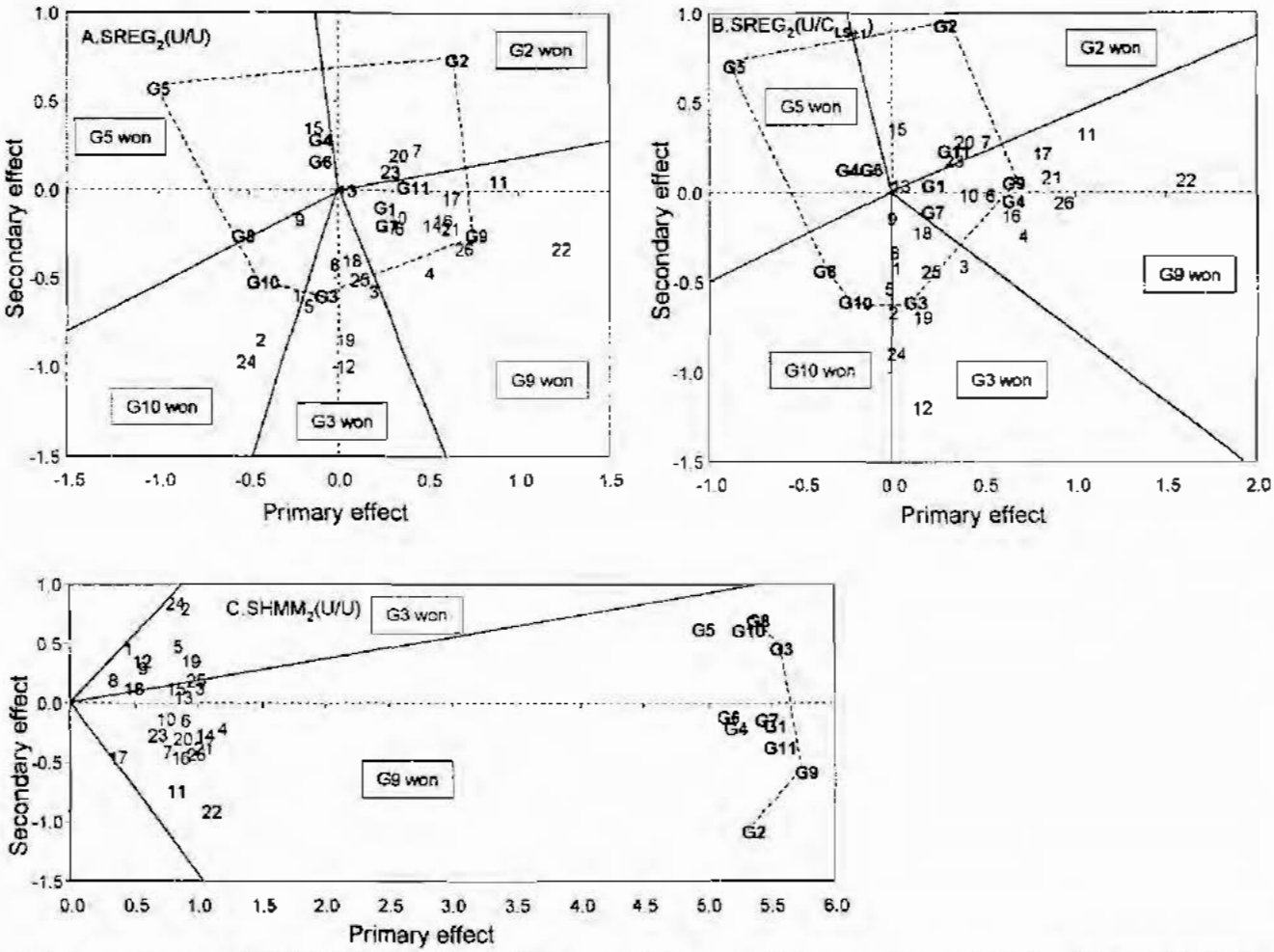

Fig. 6. Trial 2 biplots of: (A) $SREG_2(U/U)$ = shifted multiplicative model on unscaled data and unconstrained solution; (B) $SREG_2(U/C_{LS-1})$ = shifted multiplicative model on unscaled data and a constrained SVD solution; (C) $SHMM_2(U/U)$ = shifted multiplicative model on unscaled data and unconstrained solution.

nations previously described. The advantage of the biplots, however, is that sites and cultivars can be simultaneously clustered into subsets with non-COI.

The two final clusters of cultivars obtained by the fusion method clearly separated cultivars on the basis of the maize population and the year of selection and had no more than three significant COI between all possible 2 × 2 within-cluster crossover interactions (Crossa and Cornelius, 1993). Cultivars G4, G5, and G6 were selected from CIMMYT maize Population 44 in 1984, whereas Cultivars G1, G2, and G3 were selected from CIMMYT maize population 34 in 1981. Two of the three final groups of sites obtained from the empirical cell means as input data (Fig. 4) are different than those obtained from the SREG shrinkage estimates as input data (Fig. 5). But the three final groups of sites formed using the SREG shrinkage estimates, with the exception of grouping Site 8 with Sites 1, 3, and 10, agreed with subsets of sites delineated by sectors of the biplot obtained from the $SREG_2(U/U)$ model (Fig. 1A) in which Cultivars G4, G5, and G8 were the winners. Furthermore, the group of sites [7, 17, 19, and 20] tended to cluster together in most of the biplots, including those from the constrained $SREG_2$ and $SHMM_2$ models.

Trial 2

Unconstrained and Constrained $SHMM_2$ and $SREG_2$ Solutions and Their Biplots. Tests of the statistical significance of the multiplicative terms (Cornelius et al., 1992) needed to describe the variation in the Trial 2 data showed that two multiplicative components were significant ($P < 0.05$) for SREG and SHMM. The $SREG_2(U/U)$ biplot (Fig. 6A) is divided into five sectors by five winning cultivars: cultivar G2 was the predicted winner in four environments (7, 13, 20, and 23); G3 won in seven environments (3, 5, 8, 12, 18, 19, and 25); G5 won in a single environment, 15; G9 won in 10 environments (4, 6, 10, 11, 14, 16, 17, 21, 22, and 26); and G10 won in four environments (1, 2, 9, and 24). Cultivar G9 won in more environments than any other cultivar and had the highest mean yield. Cultivar G3 won in the second highest number of environments and had second highest mean yield. COI pattern is evident from opposite sectors of Fig. 6A. For example, G10 is the predicted winner at Sites 1, 2, 9, and 24 and the predicted loser at Sites 7, 20, and 23; also G2 is the worst cultivar at Sites 1, 2, and 24 and the winner at Sites 7, 20, and 23; the observed values (Table 2) confirmed these approxi-

mations. A similar COI pattern can be found between the sector where cultivar G5 is the winner versus the sector where G9 is the winner.

The $SREG_2(U/C_{LS+1})$ biplot (Fig. 6B) is virtually the same as the $SREG_2(U/U)$ biplot (Fig. 6A), except that sites with $\hat{\gamma}_{j1} < 0$ on the unconstrained solution are now forced to have $\hat{\gamma}_{j1} = 0$. No solution was obtained for the SVD non-COI constrained $SREG_2$, probably because as many as six environments had primary effects with a different sign for the primary effect than did the complementary subset consisting of 20 environments. The $SHMM_2(U/U)$ biplot (Fig. 6C) showed that all cultivars and sites have primary effects of the same sign, reflecting a complete non-COI, and that only G9 and G3 are winners, whereas the $SREG_2(U/U)$ biplot showed these as winners in only 17 of the 26 sites. The discrepancy is probably due to greater power of $SREG_2$ to detect COI.

Results of the clustering of cultivars into groups with non-COI showed two main groups {G1, G11, G4, G6, G7, G2, G9} vs. {G3, G10, G8, G5} (dendrogram not shown). These two groups are clearly separated in the three biplots (Fig. 6A–6C). The sites are clustered into three major groups {1, 24, 9, 3, 25, 8, 12, 5, 19}, {4, 6, 11, 14, 17, 26, 7, 13} and {10, 20, 23, 18, 21} and Sites 2, 16, 15, and 22 are left unclustered. The first of these site groups tended to cluster in the lower left quadrant of the $SREG_2(U/U)$ biplot (Fig. 6A), whereas the latter two groups are located toward the lower right and upper right quadrants.

In summary, the wheat data set confirmed the findings from the maize data that both $SREG_2$ and $SHMM_2$ biplots can be used to identify subsets of cultivars and sites with COI and non-COI. Since the $SREG_2$ focuses on and explains more of the cultivar main effect and the GEI, which are the sources of yield variation that are relevant to cultivar evaluation and cultivar performance based on megaenvironment identification, the $SREG_2$ biplot gives good discrimination and resolution of the cultivars and the sites. This is consistent with the conclusion of Crossa and Cornelius (1997) when comparing $SREG_1$ with $SHMM_1$ in studying COI.

CONCLUSIONS

Biplots from $SHMM_2$ and $SREG_2$ models can graphically display the interaction variation due to low level COI or non-COI (first multiplicative term) versus the interaction variation due to COI (second multiplicative term). This is accomplished if, and only if, the scores of the first singular vector of sites, $\hat{\gamma}_{j1}$, are of the same sign.

The biplots obtained using the constrained non-COI first term solutions for the $SREG_2$ and $SHMM_2$ models have the same interpretability properties as the standard biplots obtained using the unconstrained solution and give a good approximation to the patterns existing in the observed data. However, the biplot based on the unconstrained solution explains more variation and, therefore, has greater power to separate both cultivars and sites. With the constrained solution, it is possible to identify subsets of sites and cultivars with low level COI and non-COI.

Results of this study indicate that the biplots of the $SREG_2$ and $SHMM_2$ models are useful for identifying subsets of sites and cultivars with COI, low level COI, and non-COI. In general, biplots based on unscaled or scaled data gave rise to similar results. Groups of sites and cultivars with low level COI and non-COI were similar to those found when only sites (or cultivars) were clustered into non-COI groups using the SHMM and SREG clustering approach. This result confirms the benefits of using the biplots for finding simultaneous subsets of sites and cultivars with COI, low-level COI, and non-COI for breeding and agronomic purposes.

APPENDIX 1

Find $\hat{\beta}$ such that

$$\hat{\gamma}_{im} = \frac{\sum_j \hat{\alpha}_{j1}(w_{im} - \hat{\beta})}{\hat{\lambda}_1} = 0 \qquad [A1]$$

where $w_{im} = \bar{y}_{im.} - \bar{y}_{i..}$. Evidently, this should lead to

$$\hat{\beta} = \frac{\sum_j \hat{\alpha}_{j1} w_{im}}{g\sum_j \hat{\alpha}_{j1}}$$

as the solution. But if one begins the iteration with the unconstrained solution ($\hat{\beta} = 0$), then

$$\sum_j \hat{\alpha}_{j1} = 0$$

which results in division by zero. It is not known whether some other set of $\hat{\alpha}_{j1}$ values will allow an iterative solution to be computed. But

$$\hat{\alpha}_{j1} = \frac{\sum_i \hat{\gamma}_{j1}(w_{ij} - \hat{\beta})}{\hat{\lambda}_1} \qquad [A2]$$

Substitution of [A2] into [A1] gives

$$\begin{aligned}\hat{\gamma}_{im} &= \frac{\sum_j \sum_i \hat{\gamma}_{j1}(w_{ij} - \hat{\beta})(w_{im} - \hat{\beta})}{\hat{\lambda}_1^2} = 0 \\ &= \sum_j \hat{\gamma}_{j1} \sum_i (w_{ij} - \hat{\beta})(w_{im} - \hat{\beta}) \\ &= \sum_j \hat{\gamma}_{j1} [\sum_i w_{ij} w_{im} - \hat{\beta}(\sum_i w_{ij} + \sum_i w_{im}) + g\hat{\beta}^2] \\ &= \sum_j \hat{\gamma}_{j1} \sum_i w_{ij} w_{im} + g\hat{\beta}^2 \sum_j \hat{\gamma}_{j1} = 0 \qquad [A3]\end{aligned}$$

because

$$\sum_i w_{ij} = \sum_i w_{im} = 0.$$

Solving [A3] for $\hat{\beta}$ gives

$$\hat{\beta} = \pm \sqrt{\frac{-\sum_i (\sum_j \hat{\gamma}_{j1} w_{ij}) w_{im}}{g\sum_j \hat{\gamma}_{j1}}} \qquad [A4]$$

Putting $\hat{\gamma}_{im} = 0$ in [A4] gives the solution shown as Eq. [1] in the text.

APPENDIX 2

Table A1. Grain-yield rank of nine (G1–G9) maize cultivars at each of 20 test sites (Trial 1) predicted by the sites regression model ($SREG_2$) fitted to unscaled and scaled data with unconstrained solutions, constrained SVD (singular value decomposition) solutions and constrained LS (least squares) solutions and Mandel's first term solution plus one additional term ($SREG_{M+1}$).

	Site																				Site													
Cultivar	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	$SREG_2(U/U)$†																				$SREG_2(U/C_{LS+1})$																	
G1	4	8	4	8	8	8	8	1	8	2	8	8	8	8	9	8	7	8	8	7	9	8	1	8	8	5	9	1	7	1	8	6	6	8	8	8	9	8	9	8
G2	6	7	6	7	7	7	9	2	7	3	7	7	7	7	8	7	9	7	9	9	8	7	2	7	7	7	7	2	8	2	7	8	7	7	7	7	7	7	8	7
G3	5	6	5	5	5	6	6	5	5	4	5	6	6	5	5	5	6	6	6	6	7	5	3	5	5	6	5	3	5	3	5	5	5	5	5	5	6	5	6	5
G4	7	1	7	1	1	1	3	8	1	9	1	1	1	1	2	1	3	1	3	3	1	1	9	1	1	1	1	9	1	9	1	1	1	1	1	1	1	1	1	1
G5	2	4	2	2	2	4	1	9	3	6	3	4	4	3	1	3	1	4	1	1	2	3	8	3	3	3	3	8	3	8	3	3	3	3	3	3	3	3	3	3
G6	3	3	3	3	3	3	2	7	2	7	2	3	2	2	3	2	2	3	2	2	3	2	7	2	2	2	2	7	2	7	2	2	2	2	2	2	2	2	2	2
G7	8	5	8	6	6	5	7	3	6	5	6	5	5	6	7	6	8	5	7	8	6	6	4	6	6	8	6	4	6	4	6	7	8	6	6	6	5	6	5	6
G8	1	9	1	9	9	9	5	4	9	1	9	9	9	9	6	9	4	9	5	4	5	9	5	9	9	9	8	5	9	5	9	9	9	9	9	9	8	9	7	9
G9	9	2	9	4	4	2	4	6	4	8	4	2	3	4	4	4	5	2	4	5	4	4	6	4	4	4	4	6	4	6	4	4	4	4	4	4	4	4	4	4
	$SREG_2(U/C_{SVD})$																				$SREG_2(S/C_{LS+1})$																	
G1	1	7	1	8	8	7	5	2	7	1	8	8	8	7	8	8	7	8	5	6	1	8	1	8	8	6	8	1	7	1	8	7	6	8	8	8	8	8	9	8
G2	4	8	4	7	7	8	8	3	8	4	7	7	7	8	7	7	8	7	8	8	2	7	2	7	7	7	7	2	6	2	7	6	7	6	7	6	7	7	8	7
G3	2	5	2	5	5	5	4	4	5	2	5	5	5	5	5	5	5	5	4	4	3	5	3	5	5	5	5	3	5	3	5	5	5	5	5	5	5	5	6	5
G4	8	3	8	1	1	3	3	9	3	8	1	1	1	3	1	2	3	1	3	3	9	1	9	1	1	1	1	9	1	9	1	1	1	1	1	1	1	1	1	1
G5	6	2	6	3	3	2	2	7	2	5	3	3	3	2	3	3	2	3	2	2	8	3	8	3	2	3	2	8	3	8	2	3	3	3	2	3	2	3	2	2
G6	3	1	5	2	2	1	1	6	1	3	2	2	2	1	2	1	1	2	1	1	7	2	7	2	3	2	3	7	2	7	3	2	2	2	3	2	3	2	3	3
G7	7	6	7	6	6	6	7	5	6	7	6	6	6	6	6	6	6	6	7	7	5	6	5	6	6	8	6	5	8	5	6	8	8	7	6	7	6	6	5	6
G8	5	9	3	9	9	9	9	1	9	6	9	9	9	9	9	9	9	9	9	9	4	9	4	9	9	9	9	4	9	4	9	9	9	9	9	9	9	9	7	9
G9	9	4	9	4	4	4	6	8	4	9	4	4	4	4	4	4	4	4	6	5	6	4	6	4	4	4	4	4	6	4	6	4	4	4	4	4	4	4	4	4
	$SREG_2(S/U)$																				$SREG_{M+1}(U)$																	
G1	4	8	2	9	9	8	9	1	8	2	8	8	8	8	9	8	8	8	9	9	2	8	2	8	8	5	5	2	6	2	8	8	7	8	6	7	5	8	8	5
G2	6	7	3	7	8	7	8	2	7	3	7	7	7	7	8	7	7	7	8	7	4	6	6	7	7	8	7	3	8	3	7	7	8	7	8	8	8	6	7	7
G3	8	6	4	5	5	6	5	5	5	4	5	6	5	5	5	5	6	6	5	6	3	7	3	5	6	4	4	4	5	4	6	6	5	5	4	5	4	7	5	4
G4	7	1	9	1	1	1	1	9	1	9	1	1	1	1	1	1	3	1	1	2	9	1	8	2	1	3	3	9	3	9	1	1	3	1	3	3	3	1	2	3
G5	3	2	8	2	2	3	2	8	2	8	2	2	2	2	2	2	1	2	2	1	5	4	1	3	4	2	1	6	2	6	4	4	2	3	2	2	1	4	3	1
G6	2	3	7	3	3	5	3	7	3	6	3	3	3	3	3	3	2	3	3	3	7	3	4	1	2	1	2	8	1	8	2	2	1	2	1	1	2	3	1	2
G7	9	5	5	6	6	4	7	3	6	5	6	5	6	6	6	6	9	5	7	8	6	5	7	6	5	7	8	5	7	5	5	5	6	6	7	6	7	5	6	8
G8	1	9	1	8	7	9	6	4	9	1	9	9	9	9	7	9	4	9	6	5	1	9	5	9	9	9	9	1	9	1	9	9	9	9	9	9	9	9	9	9
G9	5	4	6	4	4	2	4	6	4	7	4	4	4	4	4	4	5	4	4	4	8	2	9	4	3	6	6	7	4	7	3	3	4	4	5	4	6	2	4	6
	$SREG_2(S/C_{SVD})$																				$SREG_{M+1}(S)$																	
G1	1	8	1	8	8	7	6	1	6	1	8	8	8	6	8	8	7	8	5	7	9	6	1	8	9	5	8	1	6	5	8	6	6	7	8	6	9	6	9	9
G2	4	7	4	7	7	8	7	3	8	3	7	7	7	8	7	7	8	7	6	8	5	8	4	7	8	7	7	2	8	2	7	8	8	7	7	8	8	8	8	8
G3	3	5	3	5	5	5	4	4	5	4	5	5	5	5	5	5	5	5	4	5	8	4	2	5	5	4	5	4	5	6	5	4	4	5	5	5	6	5	6	6
G4	8	1	8	1	1	2	3	9	3	9	1	1	1	3	1	1	2	1	3	2	3	3	7	2	2	3	3	8	2	7	2	3	3	3	2	2	1	3	1	2
G5	7	3	7	2	3	3	2	8	2	7	2	3	3	2	2	2	3	3	2	3	4	2	6	3	3	2	2	7	3	8	3	2	2	2	3	3	3	2	3	3
G6	2	2	2	3	2	1	1	6	1	6	3	2	2	1	3	3	1	2	1	1	7	1	5	1	1	1	1	9	1	9	1	1	1	1	1	1	2	1	2	1
G7	6	6	6	6	6	6	8	5	7	5	6	6	6	7	6	6	6	6	8	6	6	7	3	6	7	6	6	3	7	3	6	7	7	6	6	7	7	7	7	7
G8	5	9	5	9	9	9	9	2	9	2	9	9	9	9	9	9	9	9	9	9	1	9	9	9	6	9	9	5	9	1	9	9	9	9	9	9	5	9	5	5
G9	9	4	9	4	4	4	5	7	4	8	4	4	4	4	4	4	4	4	7	4	2	5	8	4	4	8	4	6	4	4	4	5	5	4	4	4	4	4	4	4

† $SREG_2(U/U)$ = sites regression model on unscaled data and unconstrained solution;
$SREG_2(U/C_{SVD})$ = sites regression model on unscaled data and constrained SVD solution;
$SREG_2(S/U)$ = sites regression model on scaled data and unconstrained solution;
$SREG_2(S/C_{SVD})$ = sites regression model on scaled data and constrained SVD solution;
$SREG_2(U/C_{LS+1})$ = sites regression model on unscaled data and constrained LS + 1 solution;
$SREG_2(S/C_{LS+1})$ = sites regression model on scaled data and constrained LS + 1 solution;
$SREG_{M+1}(U)$ = Mandel's sites regression model on unscaled data and unconstrained solution;
$SREG_{M+1}(S)$ = Mandel's sites regression model on scaled data and unconstrained solution.

Table A2. Grain-yield ranks of nine (G1–G9) maize cultivars at each of 20 test sites (Trial 1) predicted by the shifted multiplicative model ($SHMM_2$) fitted to unscaled and scaled date, with unconstrained, constrained SVD (singular value decomposition) and constrained LS (least squares) solutions.

	Site																			
Cultivar	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	$SHMM_2(U/U)$†																			
G1	2	8	1	8	8	4	9	1	8	6	8	6	8	5	8	8	8	9	9	6
G2	3	7	2	7	7	7	8	2	7	3	7	7	7	7	7	7	7	7	8	7
G3	5	6	3	5	5	5	7	4	5	4	6	5	5	6	5	5	5	6	7	5
G4	9	1	7	1	1	2	1	9	1	8	1	2	1	2	1	1	1	1	2	1
G5	8	3	4	3	3	1	4	6	2	9	3	1	2	1	3	2	2	3	6	2
G6	7	2	8	2	2	3	3	8	3	7	2	3	3	3	2	3	3	2	3	3
G7	4	5	5	6	6	8	5	5	6	2	5	8	6	8	6	6	6	5	5	8
G8	1	9	6	9	9	9	6	3	9	1	9	9	9	9	9	9	9	8	4	9
G9	6	4	9	4	4	6	2	7	4	5	4	4	4	4	4	4	4	4	1	4
	$SHMM_2(U/C_{SVD})$																			
G1	1	8	1	8	8	5	8	1	7	1	8	6	8	6	8	8	8	8	9	6
G2	3	7	3	7	7	7	7	2	8	2	7	7	7	7	7	7	7	7	7	7
G3	2	5	2	5	5	4	5	4	5	3	6	5	5	4	5	5	5	6	6	5
G4	8	1	8	1	1	3	1	9	1	9	1	2	1	3	1	1	1	1	1	3
G5	4	3	4	3	3	1	3	6	2	6	3	1	3	1	3	2	2	3	2	1
G6	5	2	5	2	2	2	2	7	3	7	2	3	2	2	2	3	3	2	3	2
G7	6	6	6	6	6	8	6	5	6	5	5	8	6	8	6	6	6	5	5	8
G8	7	9	7	9	9	9	9	3	9	4	9	9	9	9	9	9	9	9	8	9
G9	9	4	9	4	4	6	4	8	4	8	4	4	4	5	4	4	4	4	4	4
	$SHMM_2(S/U)$																			
G1	1	8	1	8	8	6	8	1	8	1	8	8	8	6	8	8	8	8	8	7
G2	2	6	2	7	7	7	6	3	7	2	7	6	7	7	7	6	6	7	7	6
G3	4	5	3	5	5	5	5	4	5	3	5	5	5	5	5	5	5	5	5	5
G4	8	1	8	1	1	2	1	9	1	8	1	1	1	2	1	1	1	1	1	1
G5	7	3	7	3	3	3	3	8	3	7	2	3	2	3	2	3	3	2	2	3
G6	6	2	5	2	2	1	2	6	2	5	3	2	3	1	3	2	2	3	3	2
G7	5	7	6	6	6	8	7	5	6	6	6	7	6	8	6	7	7	6	6	8
G8	3	9	4	9	9	9	9	2	9	4	9	9	9	9	9	9	9	9	9	9
G9	9	4	9	4	4	4	4	7	4	9	4	4	4	4	4	4	4	4	4	4
	$SHMM_2(U/C_{LS+1})$																			
G1	1	8	1	8	7	4	8	1	7	1	8	6	7	6	8	8	7	8	9	6
G2	2	7	3	7	6	7	7	2	8	2	7	7	6	7	7	7	6	7	7	7
G3	3	5	2	5	4	5	5	4	5	4	6	5	4	4	5	5	4	6	6	5
G4	8	1	8	1	1	3	1	9	1	9	1	2	1	3	1	1	1	1	1	2
G5	4	3	4	3	3	1	3	6	2	6	3	1	2	1	3	2	2	3	4	1
G6	6	2	5	2	2	2	2	7	3	7	2	3	3	2	2	3	3	2	3	3
G7	5	6	6	6	5	8	6	5	6	5	5	9	5	8	6	6	5	5	5	8
G8	7	9	7	9	8	9	9	3	9	3	9	8	8	9	9	9	8	9	8	9
G9	9	4	9	4	9	6	4	8	4	8	4	4	9	5	4	4	9	4	2	4

† $SREG_2(U/U)$ = shifted multiplicative model on unscaled data and unconstrained solution;
$SREG_2(U/C_{SVD})$ = shifted multiplicative model on unscaled data and constrained SVD solution;
$SREG_2(S/U)$ = shifted multiplicative model on scaled data and unconstrained solution;
$SHMM_2(U/C_{LS+1})$ = shifted multiplicative model on unscaled data and constrained LS + 1 solution.

REFERENCES

Abdalla, O.S., J. Crossa, and P.L. Cornelius. 1997. Results and biological interpretation of shifted multiplicative model clustering of durum wheat cultivars and test site. Crop Sci. 37:88–97.

Cornelius, P.L., and J. Crossa. 1999. Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. Crop Sci. 39:998–1009.

Cornelius, P.L., J. Crossa, and M.S. Seyedsadr. 1996. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. p. 199–234. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Cornelius, P.L., and M.S. Seyedsadr. 1997. Estimation of general linear-bilinear models for two-way tables. J. Statist. Comput. Simul. 58:287–322.

Cornelius, P.L., M. Seyedsadr, and J. Crossa. 1992. Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theor. Appl. Genet. 84:161–172.

Cornelius, P.L., D.A. Van Sanford, and M.S. Seyedsadr. 1993. Clustering cultivars into groups without rank-change interactions. Crop Sci. 33:1193–1200.

Crossa, J., and P.L. Cornelius. 1993. Recent developments in multiplicative models for cultivar trials. p. 571–577. *In* D.R. Buxton et al. (ed.) International Crop Science I. CSSA, Madison, WI.

Crossa, J., and P.L. Cornelius. 1997. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. Crop Sci. 37:406–415.

Crossa, J., P.L. Cornelius, K. Sayre, and J.I. Ortiz-Monasterio R. 1995. A shifted multiplicative model fusion method for grouping environments without cultivar rank change. Crop Sci. 35:54–62.

Crossa, J., P.L. Cornelius, and M.S. Seyedsadr. 1996. Using the shifted multiplicative model cluster methods for crossover genotype-by-environment interaction. p. 175–198. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Crossa, J., P.L. Cornelius, M. Seyedsadr, and P. Byrne. 1993. A shifted multiplicative model cluster analysis for grouping environments without genotypic rank change. Theor. Appl. Genet. 85:577–586.

Gabriel, K.R. 1971. Biplot display of multivariate matrices with application to principal components analysis. Biometrika 58:453–467.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. J. Roy. Stat. Soc. Series B. 40:186–196.

Gower, J.C., and D.J. Hand. 1996. Biplots. Chapman and Hall, UK.

SAS Institute, Inc. 1989. SAS/IML Software: usage and reference, version 6. First ed. SAS Inst., Cary, NC.

Seyedsadr, M., and P.L. Cornelius. 1993. Hypothesis testing for components of the shifted multiplicative model for a nonadditive two-way table. Comm. Stat., B. Simul. Comput. 22:1065–1078.

Yan, W., P.L. Cornelius, J. Crossa, and L.A. Hunt. 2001. Comparison of two types of GGE biplots for studying genotype by environment interaction. Crop Sci. 41:656–663.

Yan, W., L.A. Hunt, Q. Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci. 40:597–605.

Interpreting Genotype × Environment Interaction in Wheat by Partial Least Squares Regression

Mateo Vargas, José Crossa,* Ken Sayre, Matthew Reynolds, Martha E. Ramírez, and Mike Talbot

ABSTRACT

The partial least squares (PLS) regression method relates genotype × environment interaction effects (GEI) as dependent variables (Y) to external environmental (or cultivar) variables as the explanatory variables (X) in one single estimation procedure. We applied PLS regression to two wheat data sets with the objective of determining the most relevant cultivar and environmental variables that explained grain yield GEI. One data set had two field experiments, one including seven durum wheat (*Triticum turgidum* L. var. *durum*) cultivars and the other, seven bread wheat (*Triticum aestivum* L.) cultivars, both tested for 6 yr. In durum wheat cultivars, sun hours per day in December, February, and March as well as maximum temperature in March were related to the factor that explained more than 39% of GEI, while in bread wheat cultivars, minimum temperature in December and January as well as sun hours per day in January and February were the environmental variables related to the factor that explained the largest portion (>41%) of GEI. The second data set had eight bread wheat cultivars evaluated in 21 low relative humidity (RH) environments and 12 high RH environments. For both low and high RH environments, results indicated that relative performance of cultivars is influenced by differential sensitivity to minimum temperatures during the spike growth period. The PLS method was effective in detecting environmental and cultivar explanatory variables associated with factors that explained large portions of GEI.

When assessing grain yield of a set of cultivars in a multi-environment trial, changes are commonly observed in the relative yield performance of cultivars with respect to each other across sites. This differential yield response of cultivars from one environment to another is called genotype × environment interaction (GEI) and can be studied, described, and interpreted by statistical models (Crossa, 1990).

A commonly used procedure for modeling statistical interaction is a simple regression of the cultivar performance on the site mean (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966). This model can be depicted in a set of straight lines with different slopes, one for each cultivar, and the heterogeneity of slopes accounts for the interaction. Since heterogeneity of slopes generally explains only a small proportion of the complex interaction, a more elaborate model would be necessary to describe the GEI. A generalization of the regression on the site mean model is the multiplicative model also called Principal Component Analysis of the GEI or Additive Main Effect and Multiplicative Interaction (AMMI) model (Gollob, 1968; Mandel, 1971; Kempton, 1984; Gauch, 1988). Crossa et al. (1990a) investigated AMMI and other procedures for grouping environments and wheat cultivars into homogeneous subsets and determining yield stability.

The AMMI model provides more opportunity for modeling and interpreting GEI than the simple regression on the site mean model because it allows modeling the GEI in more than one dimension; however, it estimates the environmental and cultivar interaction parameters by statistics derived from the observed phenotypic data themselves. When information on external environmental variables is available (i.e., precipitation, temperature, etc.), it can be correlated to or regressed on the AMMI environmental scores so that some interpretation of the causes of grain yield GEI can be attempted. However, external environmental information cannot be used directly in the AMMI model.

When additional information is available on environment, cultivars, or both, GEI can be modeled directly by the factorial regression model (Denis, 1988; van Eeuwijk et al., 1996). Since a large number of external covariables may be modeling just noise, the most explanatory covariables may be synthesized in one covariate by the reduced rank factorial regression (van Eeuwijk et al., 1996). Also, when environmental information is available, interpretation of GEI may be possible by the principal component regression procedure that relates individual environmental variables to the principal component scores of the GEI (Aastveit and Martens, 1986). However, this approach has several problems, given that (i) it is sensitive to multicollinearity and noise and is nonparsimonious, (ii) it is not easy to relate many environmental variables to several principal component factors simultaneously, and (iii) retaining the optimal number of principal components for interpretation may be difficult (Aastveit and Martens, 1986).

To overcome some of these problems, Aastveit and Martens (1986) proposed the partial least squares (PLS) regression method as a more direct and parsimonious linear model. This method consists of relating **X** and **Y** matrices in one single estimation procedure. The **Y** matrix contains site × cultivar grain yield data as dependent variables and the **X** matrix has the external environmental variables (or external cultivar variables) as the explanatory variables. In contrast to the principal component regression approach where each component is a

M. Vargas, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática, Colegio de Postgraduados, CP 56230, Montecillo, Mexico, Universidad Autónoma Chapingo, CP 56230, Chapingo, Mexico, and International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; J. Crossa, Biometrics and Statistics Unit, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; K. Sayre and M. Reynolds, Wheat Program, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; M.E. Ramírez, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática, Colegio de Postgraduados, CP 56230, Montecillo, Mexico; M. Talbot, Biomathematics and Statistics Scotland, Univ. of Edinburgh, JCMB, Kings Buildings, Edinburgh EH93JZ, UK. *Corresponding author (JCROSSA@ALPHAC.CIMMYT.MX).

Published in Crop Sci. 38:679–689 (1998).

Abbreviations: PLS, partial least squares; GEI, genotype × environment interaction; AMMI, additive main effect and multiplicative interaction.

linear combination of the variables $x_1, x_2, ..., x_k$ only, PLS is based on the component scores using both **Y** and **X** matrices. Aastveit and Martens (1986) applied the PLS regression method to study differences in straw length of 15 barley (*Hordeum vulgare* L.) cultivars across environments over 9 yr. Talbot and Wheelwright (1989) applied the PLS regression method for explaining the GEI between nine potato (*Solanum tuberosum* L.) cultivars and 12 sites using cultivar characteristics such as disease scores and drought resistance.

The Bread Wheat Program of the International Maize and Wheat Improvement Center (CIMMYT) aims to develop widely adapted, high-yielding, stable germplasm with acceptable industrial quality and resistance to combinations of environmental stresses including drought, heat and diseases. Several studies have been done to assess GEI and yield stability of CIMMYT bread and durum wheats (Pfeiffer and Braun, 1989; Crossa et al., 1990a,b; DeLacy et al., 1994; Osman et al., 1996, 1997). However, only a few attempts have been made to explain GEI in wheat grain yield by environmental variables, genotypic variables, or both (Osman et al., 1996, 1997). In this paper, we applied PLS regression to CIMMYT bread and durum wheats in multienvironment trials with the objective of determining the most important cultivar variables, the most relevant environmental conditions, or both that influence genotype × environment interaction of grain yield.

MATERIALS AND METHODS

Partial Least Squares Regression Theory

The partial least squares approach was originally developed by Wold (1966, 1975) for systems analysis and for predicting chemical variables from spectral data. In this type of situation, the number of variables (K) is much larger than the number of observations (N), and there is high collinearity among variables. Details of the PLS theory and its similarities with principal components regression and stepwise multiple linear regression are described in Aastveit and Martens (1986). A brief explanation of PLS, where only one dependent variable (y variable) is related to many explanatory variables (X variables), is given below. Its extension to more than one dependent variable is straightforward (Aastveit and Martens, 1986).

Assume that the data for K explanatory variables are given by the matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_K)$ and data for one dependent variable is given by the vector $\mathbf{y}$. Each of the $\mathbf{x}_1 ..... \mathbf{x}_K$ and $\mathbf{y}$ vectors have N dimensions corresponding to the number of observations. The $\mathbf{y}$ vector ($N \times 1$) may represent, for example, grain yield values of N cultivars tested in a site or year (or a combination of both) and the $\mathbf{x}_k$s vectors may be measurements of other cultivar variables such as number of grains per square meter, biomass, etc. Since the PLS method is not invariant to the scale of measurement, it is assumed that variables $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$, and $\mathbf{y}$ have been centered (zero mean) and scaled (unit variance).

To break up any possible dependence among the K explanatory variables, it is more convenient to write the $\mathbf{X}$ matrix in the following bilinear form:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + ... + \mathbf{t}_M\mathbf{p}_M' + \mathbf{E}_M, \quad [1]$$

where the $\mathbf{t}_m$ ($m = 1,2,...,M$) are N-dimensional vectors called scores (also known as latent variables), the $\mathbf{p}_m$ are K-dimensional vectors called X-loadings and $\mathbf{E}_M$ is the residual matrix.

The $\mathbf{y}$ vector can be written as

$$\mathbf{y} = \mathbf{t}_1q_1 + \mathbf{t}_2q_2 + ... + \mathbf{t}_Mq_M + \mathbf{f}_M, \quad [2]$$

where $\mathbf{t}_m$ ($m = 1,2,...,M$) are the same scores as in Eq. [1] and the q_m are scalars called Y-loadings. The scores can also be called X-scores or Y-scores, depending on whether Eq. [1] or [2] is considered.

The basic idea underlying the PLS method is that the relationship between $\mathbf{X}$ and $\mathbf{y}$ is transmitted through the latent variables $\mathbf{t}_m$.

Some conditions for Eq. [1] and [2] are that the $\mathbf{t}_m$ scores should be mutually orthogonal in the space R^N or that the $\mathbf{p}_m$ loadings should also be mutually orthogonal in the space R^K. If both restrictions are imposed and if in addition orthogonality is assumed between rows/columns of $\mathbf{E}_M$, then each $\mathbf{t}_m$ is a normalized eigenvector of $\mathbf{XX}'$ and each $\mathbf{p}_m$ is a normalized eigenvector of $\mathbf{X}'\mathbf{X}$. In other words, all the vectors are essentially determined by the data matrix $\mathbf{X}$. Because of the difficulties in imposing both restrictions simultaneously, one orthogonality condition should be relaxed. This is why there are two different (but equivalent) algorithms for estimating parameters of the PLS regression, depending on whether $\mathbf{t}_m$ scores or the $\mathbf{p}_m$ loadings are considered orthogonal to each other.

In univariate PLS, the algorithm (Appendix) for representing $\mathbf{X}$ and $\mathbf{y}$ as in Eq. [1] and [2] for each $m = 1, 2, ..., M$, consists of an iterative procedure that first estimates a linear combination of the $\mathbf{X}$ variables; this gives the latent vectors (also known as factors or components). The $\mathbf{y}$ variables can be optimally predicted from that latent vector by ordinary least squares regression. A second latent vector is derived from the $\mathbf{X}$ residuals and has the capacity of optimally predicting the $\mathbf{y}$ residuals from the first step. The procedure continues until the contribution of the new latent vector is small. The number of factors (latent vectors) to be retained is determined by a cross-validation procedure (Stone, 1974), and an F test proposed by Osten (1988) is used to examine the significance of each new factor (the first, second, etc.). In this study the PLS algorithm, the cross-validation procedure, and the F test were applied by a procedure implemented in GENSTAT version 5 release 3, GENSTAT (1993).

Results of the bilinear decomposition obtained from PLS can be summarized in a graphical form similar to the biplot display of Gabriel (1971), where $\mathbf{Y}$ loadings of cultivars and $\mathbf{X}$ scores of environments are represented by vectors in a space with starting points at the origin (0,0) and end points determined by the values of the loadings and/or scores.

Data Set 1

This data set consisted of two experiments, one with seven durum wheat cultivars and the other with seven bread wheat cultivars, both tested during 6 yr (1990–1995) in Ciudad Obregon, Mexico. In each year, the experiments were arranged in a randomized complete block design with three replicates. The durum and bread wheat varieties included were a historical set released from the early 1960s to the late 1980s; the order of Numbers 1 to 7 is the order of variety releases over time (Sayre et al., 1997).

In each experiment, the grain yield GEI (kg ha^{-1}) dependent variables when using cultivar explanatory variables, were represented by the $\mathbf{Y}$ matrix of size seven × six (seven rows representing cultivars and six columns representing years). Measured at the cultivar level, the 15 explanatory variables, represented by the $\mathbf{X}$ matrix of size seven × 15 (seven rows corresponding to cultivars and 15 columns corresponding to the explanatory variables) were: days to anthesis after emergence (ANT), days to maturity after emergence (MAT), days of grainfill (GFI = MAT − ANT), plant height (cm) (PLH), above ground biomass (kg ha^{-1}) (BIO), harvest index (HID), straw yield (kg ha^{-1}) (STW), number of spikes per square

meter (NSM), number of grains per square meter (NGM), number of grains per spike (NGS), thousand kernel weight (g) (TKW), weight per tiller (g) (WTI), spike grain weight (g) (SGW), vegetative growth rate (kg ha^{-1} d^{-1}) (VGR = STW/ANT) and individual kernel growth rate (mg kernel^{-1} d^{-1}) (KGR) during the grainfill period. On the other hand, 16 environmental variables were considered: mean daily maximum temperature (°C) (MT), mean daily minimum temperature (°C) (mT), monthly total precipitation (mm) (PR), and sun hours per day (SH); they were measured during the development stage of the crop, December (D), January (J), February (F), and March (M) in each of the 6 yr (1990–1995).

In both experiments, the **Y** variable corresponds to the genotype × environment interaction matrix (residual matrix after adjusting for the genotype and environment main effects). Since the PLS procedure is not invariant to scale, both **Y** and **X** variables were centered (to mean zero) and scaled (to variance one).

Data Set 2

This data set is from the first International Heat Stress Genotype Experiments (IHSGE) (Reynolds et al., 1994) and included eight bread wheat cultivars evaluated in 33 environments (combination of sites, sowing dates, and years). The testing sites were Ciudad Obregon, Mexico, planted in December; Wad Medani, Sudan; Tlaltizapan, Mexico, planted in December; Tlaltizapan, Mexico, planted in February; Lampang, Thailand; Dharwar, India; Dinajpur, Bangladesh; Aleppo, Syria; Londrina, Brazil; Kadawa, Nigeria; Brasilia, Brazil; and Jinja, Uganda, during 1990–1994. The 33 environments were divided into two subsets, 21 low relative humidity (RH) environments and 12 high RH environments.

The grain yield (kg ha^{-1}) dependent variable **Y** matrix was of size 21 × eight (21 rows corresponding to environments and eight columns corresponding to cultivars) for the low RH environments, or size 12 × eight for the high RH environments. The 13 explanatory variables in the **X** matrix of size 21 × 13 (21 rows corresponding to environments and 13 columns corresponding to explanatory variables) for the low RH environments, or size 12 × 13 for the high RH environments, were: length of the entire growth cycle (days) (CYC), mean daily minimum temperature during the entire growth cycle (°C) (mTC), mean daily maximum temperature during the entire growth cycle (°C) (MTC), sun hours per day during the entire growth cycle (SHC), mean daily minimum temperature during the vegetative stage (mTV), mean daily maximum temperature during the vegetative stage (MTV), sun hours per day during the vegetative stage (SHV), mean daily minimum temperature during the spike growth stage (mTS), mean daily maximum temperature during the spike growth stage (MTS), sun hours per day during the spike growth stage (SHS), mean daily minimum temperature during the grainfill stage (mTG), mean daily maximum temperature during the grainfill stage (MTG), and sun hours per day during the grainfill stage (SHG).

As in the first data set, the variable **Y** corresponding to the genotype × environment interaction matrix and the **X** variables were normalized by columns in order for the mean to equal zero and the variance to equal one.

RESULTS AND DISCUSSION

Data Set 1

Explaining Genotype × Environment Interaction By Cultivar Explanatory Variables

For both durum and bread wheat experiments, the analysis of variance showed that the cultivar × year interaction for grain yield was highly significant ($P < 0.0001$). The cross-validation procedure and the F test for the number of significant factors indicated that only the first factor (latent vector) was highly significant for prediction. The predictive residual sum of squares (PRESS) for the second PLS factor was only slightly greater than for the first factor. For the durum wheat experiment, the first and second PLS factors explained 56 and 13% of the GEI, respectively, explaining jointly 69% of the GEI. For the bread wheat experiment, the first and second PLS factors explained 36 and 24% of the GEI, respectively, explaining jointly 60% of the GEI. For both crops, the relatively high percentage of GEI explained by the first PLS factor was expected since several explanatory variables were components of total grain yield.

For durum wheat, the variance of the explanatory variables number of grains per spike (NGS), harvest index (HID), spike grain weight (SGW), number of grains per square meter (NGM), individual kernel growth rate (KGR), and plant height (PLH) that was explained by the first PLS factor is large (>70%) (Table 1). These variables were closely related to a factor that made a large contribution to cultivar × year interaction and, except for kernel growth rate (KGR), they had the highest positive **X** loadings, i.e., they had high correlation with grain yield. In contrast, variability of other explanatory variables such as straw yield (STW), and number of spikes per square meter (NSM) was not explained by the first PLS factor and had values close to zero for the **X** loadings. The first PLS factor explained 15 to 65% of the variability of the remaining explanatory variables. From a biological point of view, the first PLS factor can be interpreted as the contrast between grain yield components (NGS, HID, SGW, NGM, and BIO) vs. kernel growth rate (KGR), days to anthesis after emergence (ANT), thousand kernel weight (TKW) and days to maturity after emergence (MAT).

For bread wheat, the first two factors explained

Table 1. Proportion of total variance of X variables explained by the first factor and loadings of X genotypic variables for durum wheat and bread wheat experiments of Data Set 1.

Durum wheat			Bread wheat		
Variable	% variance	X loadings	Variable	% variance	X loadings
NGS†	93.6	0.3590	ANT	76.9	0.3887
HID	90.6	0.3401	GFI	75.3	−0.3288
SGW	88.6	0.3502	BIO	71.8	0.3241
NGM	87.7	0.3408	MAT	69.1	0.3231
KGR	80.6	−0.3222	NGS	67.8	0.3816
PLH	74.4	0.3239	WTI	56.8	0.2646
BIO	65.2	0.3015	PLH	56.1	0.2166
VGR	35.0	0.2260	STW	53.7	0.1987
ANT	34.6	−0.2000	SGW	33.7	0.2716
WTI	31.4	0.2229	TKW	33.4	−0.2813
TKW	27.1	−0.1847	NSM	25.0	−0.1333
MAT	18.7	−0.1421	NGM	16.5	0.2740
GFI	14.6	0.1380	HID	15.0	−0.0257
STW	1.1	−0.0203	VGR	9.2	0.0204
NSM	0.9	−0.0507	KGR	3.9	−0.1371

† NGS = number of grains per spike, HID = harvest index, SGW = spike grain weight, NGM = number of grains per square meter, KGR = individual kernel growth rate, PLH = plant height, BIO = biomass above ground, VGR = vegetative growth rate, ANT = days to anthesis after emergence, WTI = weight per tiller, TKW = thousand kernal weight, MAT = days to maturity after emergence, GFI = days for grainfill, STW = straw yield, NSM = number of spikes per square meter.

slightly less of the variance in the GEI matrix than for durum wheat (60 vs. 69%). For five explanatory variables, days to anthesis after emergence (ANT), days for grainfill (GFI), Biomass (BIO), days to maturity after the emergence (MAT), and number of grains per spike (NGS) the first factor explained more than 67% of their variability; these five variables (except for days of grainfill) had the largest positive **X** loadings (Table 1). The first factor explained 15 to 56% of the variability of weight per tiller (WTI), plant height (PLH), straw yield (STW), spike grain weight (SGW), thousand kernel weight (TKW), number of spikes per square meter (NSM), number of grains per square meter (NGM), and harvest index (HID). These variables had intermediate positive **X** loadings except for thousand kernel weight, number of spikes per square meter and harvest index that had intermediate negative **X** loadings. The first PLS factor did not have a clear interpretation; however, it seems to have been dominated for variables related to the length of the cycle or earliness.

These results indicate that the first PLS factor explained some genotypic variables that affect GEI differently in durum wheats than they did in bread wheats. While harvest index (HID), number of grains per square meter (NGM) and individual kernel growth rate (KGR) were associated with a factor that explained a large proportion of GEI in durum wheat cultivars (90, 87.7, and 80.6%, respectively), they were not explained well in bread wheat cultivars (15, 16.5, and 3.9%, respectively). On the other hand, the variables days to anthesis after emergence (ANT), days to maturity after emergence (MAT), and days of grainfill (GFI) were not well explained by the first PLS factor for durum wheat cultivars (34.6, 18.7, and 14.6%, respectively), when compared with the first PLS factor for bread wheat cultivars (76.9, 69.1, and 75.3%, respectively). The only variables that were explained in a relatively high proportion in both crops, were number of grains per spike (NGS) and biomass (BIO) (>65%).

Table 2. Proportion of total variance of X variables explained by the first factor and loadings of X environmental variables for durum wheat and bread wheat experiments of Data Set 1.

Durum wheat			Bread wheat		
Variable	% variance	X loadings	Variable	% variance	X loadings
SHF†	82.6	−0.4169	mTJ	75.1	0.4491
MTM	69.1	0.4447	SHF	68.4	−0.4500
SHD	62.5	−0.3059	mTD	58.7	0.3285
SHM	60.4	0.4678	SHJ	53.6	−0.3267
mTD	46.1	0.2313	SHD	47.8	−0.3283
MTD	38.3	−0.3040	mTF	41.7	0.2444
mTF	28.3	0.2089	PRF	35.9	0.2674
PRM	19.8	−0.2152	MTD	34.3	−0.1610
mTJ	14.2	0.1893	PRJ	23.8	0.1610
MTJ	9.2	0.0968	SHM	11.7	0.1702
mTM	8.4	0.0122	MTM	11.7	0.1311
PRJ	6.4	−0.1300	mTM	10.9	0.1358
PRD	3.0	−0.0978	MTJ	9.4	−0.0544
MTF	2.1	−0.0227	PRM	7.3	0.0401
SHJ	1.3	−0.0871	MTF	3.2	−0.1185
PRF	0.5	0.0115	PRD	0.2	−0.0845

† SHF = sun hours in February, MTM = maximum temperature in March, SHD = sun hours in December, SHM = sun hours in March, mTD = minimum temperature in December, MTD = maximum temperature in December, mTF = minimum temperature in February, PRM = precipitation in March, mTJ = minimum temperature in January, MTJ = maximum temperature in January, mTM = minimum temperature in March, PRJ = precipitation in January, PRD = precipitation in December, MTF = maximum temperature in February, SHJ = sun hours in January, PRF = precipitation in February.

Explaining Grain Yield Variability By Environmental Explanatory Variables

Cross-validation assessment and the *F* test indicated that only one PLS factor was significant for explaining the GEI. The predictive residuals sum of squares (PRESS) for the second PLS factor was only slightly greater than for the first factor. The first factor explained 40 and 42% of the GEI in **Y** for durum wheat and bread wheat, respectively, whereas the second factor explained 26 and 20.25% of the GEI in **Y** for durum wheat and bread wheat, respectively. The first two factors explained jointly 66 and 62% of the GEI for durum and bread wheat cultivars, respectively.

For durum wheat the first PLS factor explained more than 60% of the total variability of sun hours per day in February (SHF), mean daily maximum temperature in March (MTM), sun hours per day in December (SHD) and sun hours per day in March (SHM) and had the highest relative **X** loadings (Table 2). Variability in mean daily maximum temperature in January (MTJ), minimum temperature in March (mTM), precipitation in January (PRJ), precipitation in December (PRD), maximum temperature in February (MTF), sun hours per day in January (SHJ), and precipitation in February (PRF) was not explained well by the first factor (<10%). Fifteen to 45% of the variation in mean daily minimum temperature in December (mTD), maximum temperature in December (MTD), minimum temperature in February (mTF), precipitation in March (PRM), and minimum temperature in January (mTJ) was explained by the first PLS factor. The first PLS factor can be interpreted basically as the contrast between sun hours in December, in February, and maximum temperature in December (SHD, SHF, and MTD) vs. maximum temperature and sun hours in March (MTM and SHM).

The biplot (Fig. 1a) of the first and second PLS factors for the seven durum wheat cultivars and the 6 yr shows that the first factor contrasted early released Cultivars 1 and 2 (4437 and 5188 kg ha^{-1}, respectively) with later released Cultivars 5 and 6 (7609 and 7597 kg ha^{-1}, respectively). This first factor was dominated by differences in grain yield between high yielding years 1990, 1991 and 1994 (Fig. 1a and Table 3) vs. lower yielding years 1992, 1993, and 1995. By observing both Fig. 1a and 1b simultaneously, it can be seen that the first factor also related the differences between Cultivars 1 and 2 vs. Cultivars 5 and 6 and high vs. low yielding years (Fig. 1a) with the contrast between precipitation in December, January, and March, sun hours in December, January, and February, maximum temperature in December and February (PRD, PRJ, PRM, SHD, SHJ, SHF, MTD, and MTF) (with negative **X** loadings) vs. minimum temperatures in December, January, February, and March, maximum temperatures in January, in March, and sun hours in March (mTD, mTJ, mTF, mTM, MTJ, MTM, and SHM) (with positive **X** loadings) (Fig. 1b).

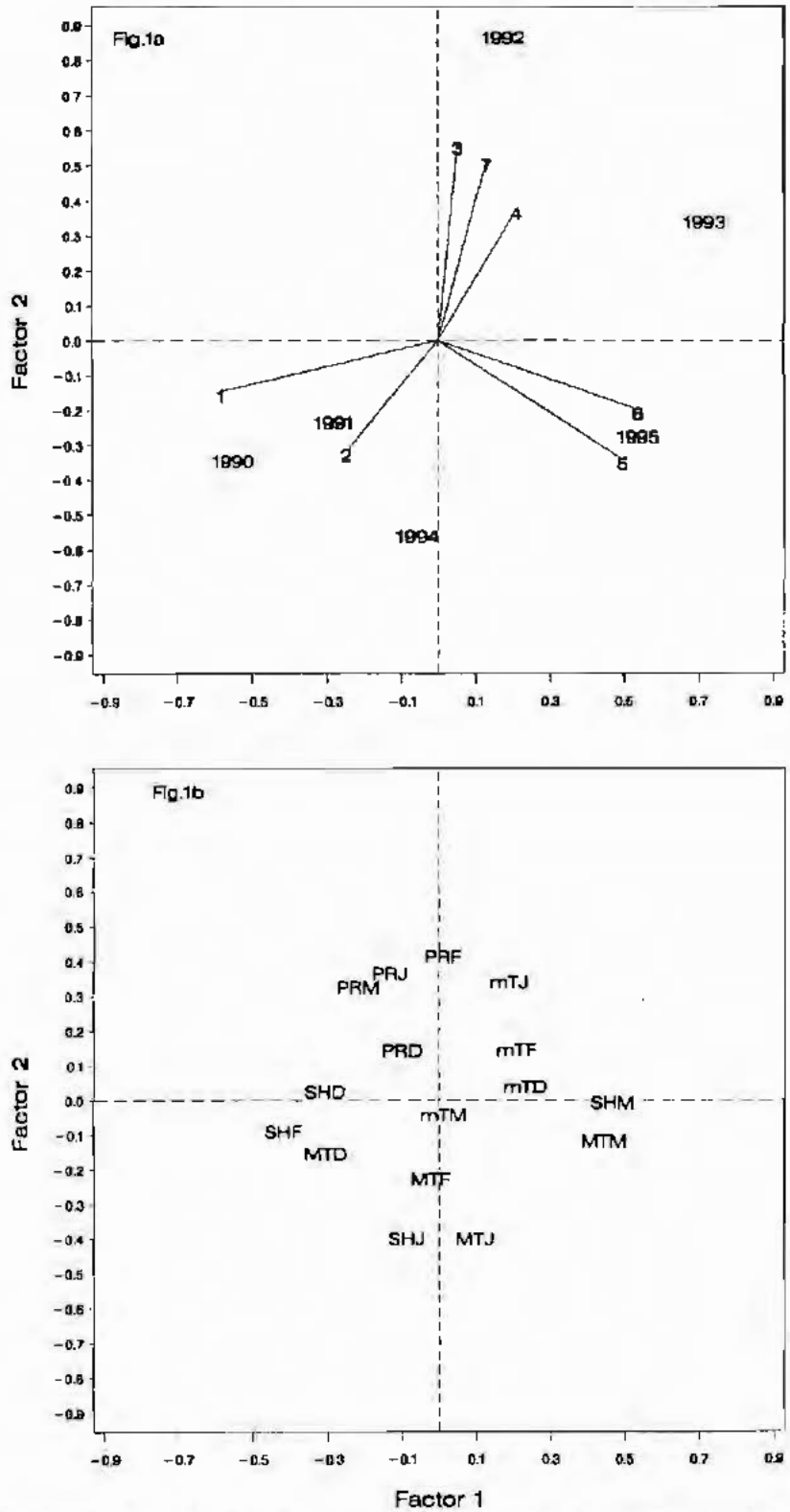

Fig. 1. (a) Biplot of the first and second PLS factors representing the X scores of years 1990, 1991, 1992, 1993, 1994, and 1995 and the Y loadings of seven durum wheat cultivars (1–7) from Data Set 1. (b) Plot of the first two PLS factors representing the X loadings of 16 environmental variables measured in 6 yr when seven durum wheat cultivars were evaluated (Data Set 1). Environmental variables are denoted as mTD = minimum temperature in December, mTJ = minimum temperature in January, mTF = minimum temperature in February, mTM = minimum temperature in March, MTD = maximum temperature in December, MTJ = maximum temperature in January, MTF = maximum temperature in February, MTM = maximum temperature in March, PRD = precipitation in December, PRJ = precipitation in January, PRF = precipitation in February, PRM = precipitation in March, SHD = sun hours in December, SHJ = sun hours in January, SHF = sun hours in February, SHM = sun hours in March.

In general, the highest yielding years (1990, 1991, and 1994) had more sun hours and lower minimum temperatures than the lowest yielding years (1992, 1993, 1995) (Table 3); these environmental conditions favored Cultivars 1 and 2 more. Grain yield had high positive correlations with sun hours in January ($r = 0.70$) and February ($r = 0.69$) and maximum temperatures in December ($r = 0.76$) and February ($r = 0.55$), and it had negative correlations with minimum temperatures. Early released Cultivars 1 and 2 had a positive interaction with years 1990, 1991, and 1994 because they are all located in the same quadrant (lower left) of Fig. 1a. This interaction seems to be associated with sun hours in January and February (SHJ and SHF) and higher maximum temperatures in December and February (MTD and MTF) (Fig. 1b). The negative interaction between early released Cultivars 1 aud 2 with years 1992 and 1993 (they are in opposite quadrants) seems to be due to lower minimum temperatures in December, January, and February (mTD, mTJ, and mTF) as well as more precipitation in February (PRF). However, these environmental conditions during the 1992 and 1993 growth cycle favored intermediate released Cultivars 3 and 4 and later released Cultivar 7. Later released Cultivars 5 and 6 showed positive interactions with year 1995 because of higher maximum temperatures in January and March (MTJ and MTM).

For bread wheat, total variability in minimum temperature in December and January (mTD and mTJ) and sun hours per day in January and February (SHJ and SHF) was explained well by the first PLS factor (>53%) (Table 2); mTD and mTJ had high positive loadings, whereas SHJ and SHF had high negative loadings. The first factor explained 24 to 48% of the variability of sun hours per day in December (SHD), minimum temperature in February (mTF), precipitation in February (PRF), maximum temperature in December (MTD), and precipitation in January (PRJ) (with intermediate absolute loadings). For durum wheat and bread wheat, variability of maximum temperature in January and February (MTJ and MTF), minimum temperature in March (mTM), and precipitation in December (PRD) was not explained well by the first factor (<11%). For bread wheat, the first factor did not explain as much variability in maximum temperature in March (MTM) and sun hours per day in March (SHM) as it did for durum wheat.

The **X** scores and **Y** loadings for the first and second PLS factors obtained for the bread wheat cultivars followed patterns similar to those for the durum wheats; that is, the first factor contrasted high vs. low yielding years, and precipitation and minimum temperatures vs. maximum temperatures and sun hours (not shown). Grain yield of the earlier Bread Wheat Releases 2, 3, 4, and 5 was more favored by sun hours in December, January and February and by maximum temperatures in December, January and February 1990 , 1991, and 1994, whereas yield of later Bread Wheat Releases 6 and 7 was more positively associated with maximum temperatures in March and sun hours per day in March 1993.

Table 3. Mean grain yield (Yield), minimum temperature (mT), maximum temperature (MT), precipitation (PR) and sun hours (SH) for each of the years (Data Set 1) when seven durum wheat cultivars were tested.

Year	Yield	mT	MT	PR	SH
	kg ha^{-1}	——— °C ———		mm	hours
1990	7509.4†	4.42	25.15	13.75	8.33
1994	7481.2	7.20	26.02	0.50	7.44
1991	6987.0	6.75	25.37	27.25	8.09
1995	6243.8	7.77	25.97	21.50	7.80
1992	5978.5	7.97	24.20	79.75	7.06
1993	5742.6	8.17	25.25	22.25	7.15

† Grain yields in years 1990, 1991, and 1994 were significantly different from grain yields in years 1992, 1993, and 1995 ($P < 0.05$).

In summary, sun hours per day in February (SHF) was associated with a factor that explained a large proportion of GEI in both crops, and maximum temperature in January and in February (MTJ and MTF), minimum temperature in March (mTM) and precipitation in December (PRO) were associated with a factor that explained a small proportion of GEI in both crops. However, other environmental variables such as maximum temperature in March (MTM) and sun hours per day in March (SHM) were associated with a factor that explained more GEI in durum wheats than in bread wheats. On the other hand, minimum temperature in January (mTJ) and sun hours per day in January (SHJ) were explained well by the first PLS factor in bread wheats but not in durum wheats. It is apparent that yields of the earlier released durum and bread cultivars were more favored than yields of later released cultivars in 1990, 1991, and 1994 because of sunnier weather in December, January and February and lower minimum temperatures.

Data Set 2

Explaining Genotype × Environment Interaction By Environmental Explanatory Variables

For the 21 low relative humidity (RH) environments and the 12 high RH environments included in the IHSGE, the cross-validation assessment and the F test indicated that only one factor was highly significant for explaining GEI. The second PLS factor had a predictive residual sum of squares (PRESS) slightly greater than for the first factor. For the low RH environments, the first and second factors explained 27.4 and 8.2% of the GEI, respectively, and for the high RH environments, the first and second factors explained 18.75 and 11.86% of the GEI, respectively.

For the low RH environments, the first PLS factor explained a large proportion of the total variability of minimum temperature during the spike growth stage (mTS) (84.4%), minimum temperature during the entire growth cycle (mTC) (83.9%), and minimum temperature during the grainfill stage (mTG) (68.9%) (Table 4). These variables had the highest positive **X** loadings. Other environmental variables such as length of the entire growth cycle (CYC) and sun hours per day during the grainfill stage (SHG) (with high negative loadings) and maximum temperature during the spike growth stage (MTS) were also explained well by the first PLS

Table 4. Proportion of total variance of X variables explained by the first factor and loadings of X environmental variables for first International Heat Stress Genotype Experiments (Dataset 2).

Low relative humidity environment			High relative humidity environment		
Variable	% variance	X loadings	Variable	% variance	X loadings
mTS†	84.4	0.4274	mTS	85.4	0.4962
mTC	83.9	0.3969	mTC	81.5	0.4806
mTG	68.9	0.3733	CYC	50.1	−0.4369
CYC	52.3	−0.2921	SHG	44.4	−0.1812
SHG	51.1	−0.4311	mTG	44.1	0.2704
MTS	45.2	0.2332	SHS	41.3	−0.2343
mTV	35.3	0.2283	MTV	40.5	0.1788
MTC	29.5	0.1516	mTV	37.6	0.2942
MTG	14.5	0.1006	MTS	28.5	0.1227
SHC	9.9	−0.2564	SHC	24.4	−0.1742
MTV	8.9	0.0545	MTC	11.2	0.0092
SHS	2.7	−0.1841	SHV	3.3	−0.0839
SHV	0.0	−0.0840	MTG	2.2	0.0037

† mTS = minimum temperature during the spike growth stage, mTC = minimum temperature during the entire growth cycle, mTG = minimum temperature during the grainfill stage, CYC = length of the entire growth cycle, SHG = sun hours per day during the grainfill stage, MTS = maximum temperature during the spike growth stage, mTV = minimum temperature during the vegetative stage, MTC = maximum temperature during the entire growth cycle, MTG = maximum temperature during the grainfill stage, SHC = sun hours per day during the entire growth cycle, MTV = maximum temperature during the vegetative stage, SHS = sun hours per day during the spike growth stage, SHV = sun hours per day during the vegetative stage.

factor (>45%). On the other hand, for sun hours per day during the vegetative stage, the spike growth stage and the entire growth cycle (SHV, SHS, and SHC), and maximum temperature during the vegetative stage (MTV) the first factor made a negligible contribution in explaining their variabilities (<10%) (Table 4).

The biplot of the first and second factors for the eight wheat cultivars evaluated in 21 low RH environments showed that the first factor is primarily a contrast between the seven highest yielding low RH environments vs. the intermediate and low yielding low RH environments (Fig. 2a, Table 5). This first factor clearly separates sun hours (SHS, SHV, SHC, and SHG) and length of the entire growth cycle (CYC) from minimum and maximum temperatures measured during the different growth stages (Fig. 2b). It can be observed that environments with high maximum and minimum temperatures are contrasted with environments with high sun hours and a long cycle (Figs. 2a and 2b and Table 5). Therefore, the first PLS factor showed that high yielding environments had more sun hours and longer growing cycle than intermediate and low yielding environments. On the other hand, intermediate and low yielding environment had lower minimum and higher maximum temperatures than high yielding environments. Concerning the cultivars, the first factor discriminated Group 1 Cultivars 2, 3, 5, and 8 from Group 2 Cultivars 1, 4, 6, and 7 (Fig. 2a).

Group 1 Cultivars 2, 5, and 8 are concentrated in the upper left quadrant of the biplot and had positive yield interaction with environments such as December planting date in Tlaltizapan, Years 1 and 3 (TLD1 and TLD3), Sudan, Years 2 and 3 (SUD2 and SUD3), and India, Dharwar, Year 2 (IND2) (Fig. 2a). The environmental variables that seemed to be positively affecting the interaction of these cultivars in those environments were sun hours during the entire growth cycle, vegetative stage, spike growth stage, and grainfill stage (SHC, SHV, SHS, and SHG) (Fig. 2b). Grain yield had positive correlations with sun hours (data not shown). Cultivar 3 had positive interactions with environments TLD2, OBD1, and IN14 because of long cycles (CYC), 99 d, 125 d, and 93 d, respectively (Table 5). On the other hand, Group 2 Cultivars 1, 4, and 6 showed positive yield interaction with environments NIG2, NIG3, NIG4, TLF2, SYR1, and SYR2 because, in general, these environments had higher minimum temperatures than most of the others (Table 5).

For the high RH environments, the results were similar to that for the low RH environments. The first factor explained 85.4 and 81.5% of the variability of minimum temperature during the spike growth stage and entire growth cycle (mTS and mTC) and both had large positive loadings (Table 4). For length of the entire growth cycle (CYC) (with high negative loading) the first PLS factor explained 50.1% of its total variability. On the other hand, variability of maximum temperature during the grainfill stage (MTG) and sun hours per day during the vegetative stage (SHV) were not explained well by the first PLS factor (<4%). For the remaining variables the first PLS factor explained a intermediate proportion (11.2–44.4%) of their total variability.

The **X** scores and **Y** loadings for the first and second PLS factors obtained for the eight bread wheat cultivars and the 12 high RH environments followed patterns similar to those for the low RH environments, the first PLS factor contrasted seven high vs. five low yielding environments and minimum and maximum temperatures vs. sun hours during the different growth stages and total length of the entire growth cycle (biplots not shown).

In summary, for both low and high RH environments, the greatest proportion of total variability of minimum temperatures was explained by the first factor. These data indicated that relative performance of cultivars was strongly influenced by differential sensitivity to minimum temperatures that varied from 10.5°C to over 16°C during the spike growth period, depending on the environment. Possible mechanisms are genetic differences in temperature sensitivity to respiratory carbon loss, as well as the effect of temperature on rate of spike development. In both cases, cooler temperatures would be favorable to spike growth, which determines yield potential. Similarly, genetic variability for this temperature dependant process would explain the effect of maximum temperatures on yield variability, though the effect is weaker. When comparing high and low RH environments, number of sun hours tended to be more critical in determining relative performance in the high RH environments than the low RH ones. This was probably because radiation levels were generally lower and, hence, more of a limiting factor in high RH environments.

CONCLUSIONS

Results for Data Set 1 showed that while number of grains per spike and biomass were associated with a

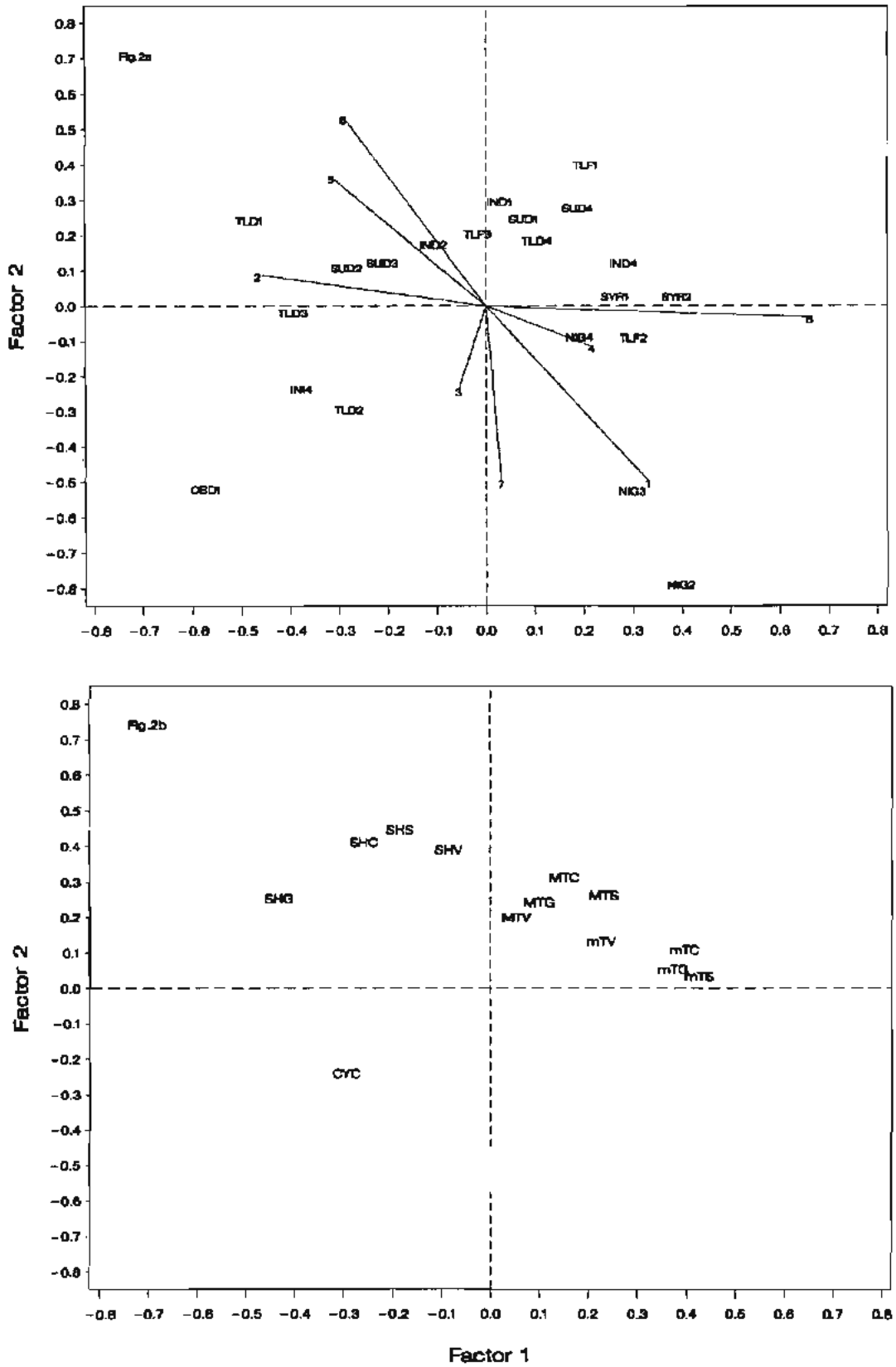

Fig. 2. Legend on next page.

Table 5. Mean grain yield (Yield), length of the entire growth cycle (CYC), minimum temperature during the entire growth cycle (mTC), maximum temperature during the entire growth cycle (MTC) and sun hours during the entire growth cycle (SHC) for 21 low relative humidity environments included in the first International Heat Stress Genotype Experiments (Data Set 2).

Environment	Yield	CYC	mTC	MTC	SHC
	kg ha^{-1}	days	°C		hours
OBD1†	5740.5	125	10.30	25.4	8.40
SUD2	5154.7	99	13.10	32.0	9.50
INI4	4933.4	93	11.10	27.7	9.00
TLD3	4534.0	106	10.88	31.5	9.20
TLD1	4172.0	99	10.50	32.2	10.00
TLD2	4084.4	99	11.30	30.5	8.40
IND2	4022.4	87	13.40	33.0	9.50
SUD1	3971.4	83	16.20	35.0	9.50
TLD4	3613.8	90	14.60	34.8	9.30
IND4	3563.4	80	17.70	33.5	9.20
TLF3	3458.6	82	13.10	34.8	9.30
SUD3	3267.0	92	13.20	32.5	9.50
IND1	3190.7	82	15.40	33.6	9.80
TLF1	2589.5	78	15.40	36.1	10.00
NIG4	2511.5	81	15.40	32.2	8.70
TLF2	2497.7	83	15.60	34.1	8.50
SUD4	2200.7	80	16.90	35.6	9.50
NIG3	1828.3	97	15.50	31.9	7.19
NIG2	1718.8	89	15.20	29.6	6.60
SYR1	1590.0	86	17.50	32.9	9.00
SYR2	864.7	88	19.00	33.0	9.00

† OBD1: Ciudad Obregon, December, year 1; SUD2: Sudan, Year 2; INI4: India, Indore, Year 4; TLD3: Tlaltizapan, December, Year 3; TLD1: Tlaltizapan, December, Year 1; TLD2: Tlaltizapan, December, Year 2; IND2: India, Dharwar, Year 2; SUD1: Sudan, Year 1; TLD4: Tlaltizapan, December, Year 4; IND4: India, Dharwar, Year 4; TLF3: Tlaltizapan, February, Year 3; SUD3: Sudan, Year 3; IND1: India, Dharwar, Year 1; TLF1: Tlaltizapan, February, Year 1; NIG4: Nigeria, Year 4; TLF2: Tlaltizapan, February, Year 2; SUD4: Sudan, Year 4; NIG3: Nigeria, Year 3; NIG2: Nigeria, Year 2; SYR1: Syria, Year 1; SYR2: Syria, Year 2.

factor that explained a large proportion of the variability on GEI in both crops (>65 and >67% for durum and bread wheats, respectively), harvest index, number of grains per square meter, and individual kernel growth rate were explained well by that factor only in durum wheats (>80%). Conversely, days to anthesis after emergence, days of grainfill and days to maturity after emergence were associated with a factor that explained GEI in bread wheat (>69%) but not in durum wheats. In relation to environmental variables, sun hours per day in February was associated with a factor that explained a large proportion of the GEI in both crops. Sun hours in December, February and March as well as maximum temperature in March were related to a factor that explained a large proportion of GEI in durum wheats. Minimum temperature in December and January as well as sun hours in January and February were associated with a PLS factor that explained a larger proportion of the GEI in bread wheats. Results indicated that grain yields of earlier released wheat cultivars (durum and bread) were more favored than the yields of later released cultivars in 1990 and 1991 because of sunnier weather during January and February, maximum temperatures in December and February, and lower minimum temperatures. Grain yields of later released durum and bread wheat cultivars were more positively affected by maximum temperatures in January and March and sun hours in March of 1993 and 1995.

Results for Data Set 2 indicated that minimum temperature during the spike growth stage, entire growth cycle, and the length of the entire growth cycle were correlated to a PLS factor that explained most of the GEI in low and high RH environments. For both low and high RH environments, results indicated that relative performance of cultivars was strongly influenced by differential sensitivity to minimum and maximum temperatures during the different growth stages. Sunnier weather and the longer cycle in high yielding environments favored some cultivars; lower minimum and higher maximum temperatures in low yielding environments favored another group of cultivars.

Results of this study indicated that the PLS method was effective in reducing information existing in four complex multivariate data sets. It effectively detected environmental and cultivar explanatory variables associated with factors that explained large proportions of GEI. In these data sets, several genotypic variables such as yield components were highly correlated; however, the PLS method deals appropriately with this problem. Furthermore, the cross-validation procedure and the F test are useful tools for determining the optimal number of significant components (factors) that are required for explaining GEI. More research is needed for comparing the PLS method with other statistical models such as the factorial regression model.

ACKNOWLEDGMENTS

The authors are thankful to researchers in national programs for conducting the experiments analyzed in this study and to Alma McNab for editing the manuscript. Thanks are due to the Associate Editor and three anonymous reviewers for their comments and suggestions that significantly improved the quality of the manuscript.

Fig. 2. (a) Biplot of the first and second PLS factors representing the X scores of 21 low relative humidity environments and the Y loadings of eight wheat cultivars (1–8) from Data Set 2. Environments are denoted as OBD1 = Ciudad Obregon, December, Year 1; SUD1 = Sudan, Year 1; SUD2 = Sudan, Year 2; SUD3 = Sudan, Year 3; SUD4 = Sudan, Year 4; NIG2 = Nigeria, Year 2; NIG3 = Nigeria, Year 3; NIG4 = Nigeria, Year 4; SYR1 = Syria, Year 1; SYR2 = Syria, Year 2; INI4 = India, Indore, Year 4; IND1 = India, Dharwar, Year 1; IND2 = India, Dharwar, Year 2; IND4 = India, Dharwar, Year 4; TLD1 = Tlaltizapan, December, Year 1; TLD2 = Tlaltizapan, December, Year 2; TLD3 = Tlaltizapan, December, Year 3; TLD4 = Tlaltizapan, December, Year 4; TLF1 = Tlaltizapan, February, Year 1; TLF2 = Tlaltizapan, February, Year 2; TLF3 = Tlaltizapan, February, Year 3; (b) Plot of the first two PLS factors representing the X loadings of 13 environmental variables measured in 21 low relative humidity environments where 8 wheat cultivars were evaluated (Data Set 2). Environmental variables are denoted as CYC = length of the entire growth cycle, mTC = minimum temperature during the entire growth cycle, MTC = maximum temperature during the entire growth cycle, SHC = sun hours per day during the entire growth cycle, mTV = minimum temperature during the vegetative stage, MTV = maximum temperature during the vegetative stage, SHV = sun hours per day during the vegetative stage, mTS = minimum temperature during the spike growth stage, MTS = maximum temperature during the spike growth stage, SHS = sun hours per day during the spike growth stage, mTG = minimum temperature during the grainfill stage, MTG = maximum temperature during the grainfill stage and SHG = sun hours per day during the grainfill stage.

APPENDIX

BRIEF DESCRIPTION OF THE UNIVARIATE PARTIAL LEAST SQUARES REGRESSION ALGORITHM

The iterative algorithm (Helland, 1988) for representing **X** and **y** as in Eq. [1] and [2] (for each value of m) consists of the following steps.

Step 1. Write

$$\mathbf{E}_0 = \mathbf{X}$$

and

$$\mathbf{f}_0 = \mathbf{y} \qquad [3]$$

Step 2. Find $\mathbf{t}_m$, $\mathbf{p}_m$, and q_m by induction.

The basic point now is that each $\mathbf{t}_m$ is determined as a linear combination of the **X** residuals obtained in the previous step. In particular for $m = 1$ one wants

$$\mathbf{t}_1 = \Sigma_{k=1,K}\, \mathbf{x}_k w_{k1} = \mathbf{X}\mathbf{w}_1 \qquad [4]$$

where $\mathbf{w}_1$ is a k-dimensional weighting vector.

Because it is desired that $\mathbf{t}_1$ should be highly correlated with **y**, it is reasonable to make each w_{k1} component proportional to the covariance between $\mathbf{x}_k$ and **y**. This is accomplished by taking $w_{k1} = \mathbf{x}_k'\mathbf{y}$, that is:

$$\mathbf{w}_1 = \mathbf{X}'\mathbf{y} \qquad [5]$$

Then by Eq. [3], the Eq. [4] and [5] become

$$\mathbf{t}_1 = \mathbf{E}_0\mathbf{w}_1$$

and

$$\mathbf{w}_1 = \mathbf{E}_0'\mathbf{f}_0.$$

Therefore, for general m we have:

$$\mathbf{w}_m = \mathbf{E}_{m-1}'\mathbf{f}_{m-1} \qquad [6]$$

and

$$\mathbf{t}_m = \mathbf{E}_{m-1}\mathbf{w}_m \qquad [7]$$

Step 3. Find the $\mathbf{p}_m$ and q_m values from the best possible fit of Eq. [1] and [2].

For $m = 1$ the best fit to

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{f}_1$$

is given by the regression coefficient

$$q_1 = (\mathbf{y}'\mathbf{t}_1)/(\mathbf{t}_1'\mathbf{t}_1),$$

when $\mathbf{f}_0 = \mathbf{y}$.

$$q_1 = (\mathbf{f}_0'\mathbf{t}_1)/(\mathbf{t}_1'\mathbf{t}_1).$$

In general for any value of m

$$q_m = (\mathbf{f}_{m-1}'\mathbf{t}_m)/(\mathbf{t}_m'\mathbf{t}_m) \qquad [8]$$

Similarly, the best fit to

$$\mathbf{x}_k = \mathbf{t}_1 p_{k1} + \mathbf{e}_{k1}$$

is given by

$$p_{k1} = (\mathbf{x}_k'\mathbf{t}_1)/(\mathbf{t}_1'\mathbf{t}_1), \qquad (k = 1,2,\ldots,K)$$

or

$$\mathbf{p}_1 = (\mathbf{X}'\mathbf{t}_1)/(\mathbf{t}_1'\mathbf{t}_1).$$

By $\mathbf{E}_0 = \mathbf{X}$,

$$\mathbf{p}_1 = (\mathbf{E}_0'\mathbf{t}_1)/(\mathbf{t}_1'\mathbf{t}_1)$$

and in general for any value of m

$$\mathbf{p}_m = (\mathbf{E}_{m-1}'\mathbf{t}_m)/(\mathbf{t}_m'\mathbf{t}_m) \qquad [9]$$

Step 4. By Eq. [3] and for $m = 1$ Eq. [1] and [2] become

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{E}_1$$

and

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{f}_1, \qquad [10]$$

respectively, from where $\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1'$ and $\mathbf{f}_1 = \mathbf{y} - \mathbf{t}_1 q_1$. After substituting $\mathbf{E}_0$ and $\mathbf{f}_0$ in Eq. [10]

$$\mathbf{E}_1 = \mathbf{E}_0 - \mathbf{t}_1\mathbf{p}_1$$

and

$$\mathbf{f}_1 = \mathbf{f}_0 - \mathbf{t}_1 q_1.$$

In general, for any value of m

$$\mathbf{E}_m = \mathbf{E}_{m-1} - \mathbf{t}_m\mathbf{p}_m'$$

and

$$\mathbf{f}_m = \mathbf{f}_{m-1} - \mathbf{t}_m q_m. \qquad [11]$$

Step 5. Repeat Steps 2 to 4 until the contribution of the new factor is small.

PREDICTION

To obtain the prediction of new values consider that $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0K})'$, a row vector in the **X** matrix, is a set of X-measurements in a new unit and define $\mathbf{e}_0 = \mathbf{x}_0 - \bar{\mathbf{x}}$ with $\bar{\mathbf{x}} = (\bar{x}_1,\ldots,\bar{x}_k)'$. Then the new scores and residuals are obtained as

$$\mathbf{t}_{m0} = \mathbf{e}_{m-1}'\mathbf{w}_m$$

and

$$\mathbf{e}_m = \mathbf{e}_{m-1} - \mathbf{t}_{m0}\mathbf{p}_m$$

Therefore, the corresponding value y_0 is predicted in step m by

$$\hat{y}_{m0} = \bar{y} + \Sigma_{m=1,M}\, \mathbf{t}_{m0} q_m$$
$$= \bar{y} + \Sigma_{m=1,M}\, \mathbf{t}_{m0}(\mathbf{t}_m'\mathbf{t}_m)^{-1}\, \mathbf{t}_m'\mathbf{y}$$

MULTIVARIATE PARTIAL LEAST SQUARES REGRESSION

The method can be used for multivariate as well as univariate regression, so there may be several dependent variables given by the matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L]$, say. To form a relation between the **Y** variables and explanatory variables $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$ similar to that of Eq. [1] and [2], the **X** and **Y** matrices can be written as

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \ldots + \mathbf{t}_M\mathbf{p}_M' + \mathbf{E}_M$$
$$= \mathbf{T}\,\mathbf{P} + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{t}_1\mathbf{q}_1' + \mathbf{t}_2\mathbf{q}_2' + \ldots + \mathbf{t}_M\mathbf{q}_M' + \mathbf{F}_M$$
$$= \mathbf{T}\,\mathbf{Q} + \mathbf{F}$$

Where the $\mathbf{q}_m$ are now L-dimensional vectors called Y-loadings and $\mathbf{F}_M$ is the residual matrix.

Aastveit and Martens (1986) gives details of the two differ-

ent multivariate PLS algorithms. One algorithm yields non orthogonal scores **T** and therefore requires a multiple linear regression stage to estimate Q. The other algorithm contains an extra step in order to orthogonalize the scores **T** and thereby simplify the estimation of Q. The two algorithms yield the same final prediction results. The orthogonalized algorithm yields a set of orthonormal loading vectors **W** and a set of nonorthogonal loading vectors **P**. The advantage of this orthogonalized algorithm version is that the parameters can be estimated for each factor separately, since both **W** and **T** are orthogonal. Hence, no matrix inversion is required. GENSTAT procedure uses the orthogonalized algorithm.

REFERENCES

Aastveit, H., and H. Martens. 1986. ANOVA interactions interpreted by partial least squares regression. Biometrics 42:829–844.

Crossa, J. 1990. Statistical analyses of multilocation trials. Adv. Agron. 44:55–85.

Crossa, J., P.N. Fox, W.H. Pfeiffer, S. Rajaram, and H.G. Gauch, Jr. 1990a. AMMI adjustment for statistical analysis of an international wheat yield trial. Theor. Appl. Genet. 81:27–37.

Crossa, J., W.H. Pfeiffer, P.N. Fox, and S. Rajaram. 1990b. Multivariate analysis for classifying sites: Application to an international wheat yield trial. *In* M.S. Kang (ed.) Genotype-by-environment interaction in plant breeding. Louisiana State University, Baton Rouge.

DeLacy, I.H., P.N. Fox, J.D. Corbett, J. Crossa, S. Rajaram, R.A. Fischer, and M. van Ginkel. 1994. Long-term association of locations for testing spring wheat. Euphytica 72:95–106.

Denis, J-B. 1988. Two-way analysis using covariates. Statistics 19: 123–132.

Eberhart, S.A., and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36–40.

Finlay, K.W., and G.N. Wilkinson. 1963. The analysis of adaptation in a plant breeding programme. Aust. J. Agric. Res. 14:742–754.

Gabriel, K.R. 1971. Biplot display of multivariate matrices with application to principal component analysis. Biometrika 58:453–467.

Gauch, H.G., Jr. 1988. Model selection and validation for yield trials with interaction. Biometrics 44:705–715.

GENSTAT. 1993. Genstat 5 release 3, reference manual. Clarendon Press, Oxford, UK.

Gollob, H.F. 1968. A statistical model which combines features of factor analysis and analysis of variance techniques. Psychometrika 33:73–115.

Helland, I.S. 1988. On the structure of partial least squares regression. Commun. Statist. Simula. 17(2):581–607.

Kempton, R.A. 1984. The use of biplot in interpreting variety by environment interactions. J. Agric. Sci. (Cambridge) 103:123–135.

Mandel, J. 1971. A new analysis of variance model for non-additive data. Technometrics 13:1–18.

Osman, S.A., J. Crossa, E. Autrique, and I.H. DeLacy. 1996. Relationships among International Testing Sites of Spring Durum Wheat. Crop Sci. 36:33–40.

Osman, S.A., J. Crossa, and P.L. Cornelius. 1997. Results and biological interpretation of shifted multiplicative model clustering of durum wheat cultivars and test site. Crop Sci. 37:88–97.

Osten, D.W. 1988. Selection of optimal regression models via cross-validation. J. Chemometrics 2:39–48.

Pfieffer, W.H., and H.J. Braun. 1989. Yield stability in bread wheat. *In* J.R. Henderson and P.B. Hazell (ed.) Variability in grain yield: Implications for agricultural research and policy developing countries. The John Hopkins University Press, Baltimore.

Reynolds, M.P., M. Balota, M.I.B. Delgado, I. Amani, and R.A. Fisher. 1994. Physiological and morphological traits associated with spring wheat yield under hot, irrigated conditions. Aust. J. Plant Physiol. 21:717–730.

Sayre, K.D., S. Rajaram, and R.A. Fisher. 1997. Yield potential progress in short bread wheats in northwest Mexico. Crop Sci. 37:36–42.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc., Ser. B, 36:111–147.

Talbot, M., and A.V. Wheelwright. 1989. The analysis of genotype × environment interactions by partial least squares regression. Biuletyn Oceny Odmian, Zeszyt 21/22:19–25.

van Eeuwijk, F.A., J-B. Denis, and M.S. Kang.1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. p. 15–49. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. p. 391–420. *In* P.R. Krishnaiah (ed.) Multivariate analysis. Academic Press, New York.

Wold, H. 1975. Soft modelling of latent variables: The nonlinear interactive partial least squares approach. p. 114–142. *In* Gani (ed.) Perspectives in probability and statistics: Papers in honour of M.S. Bartlett on the occasion of his sixty-fifth birthday. Academic Press, London.

Yates, F., and W.G. Cochran. 1938. The analysis of groups of experiments. J. Agric. Sci. (Cambridge) 28:556–580.

Relationships among Bread Wheat International Yield Testing Locations in Dry Areas

Richard M. Trethowan,* Jose Crossa, Maarten van Ginkel, and Sanjaya Rajaram

ABSTRACT

Understanding the relationship among yield testing locations is important if plant breeders are to target germplasm better to different production environments or regions. To examine the relationship among international drought prone test sites, yield data from 122 locations, sown during a 6-yr period in CIMMYT's Semi-Arid Wheat Yield Trial (SAWYT) were analyzed. The shifted multiplicative model (SHMM) was used to group locations within each year and pattern analysis was employed to group those sites across years. Sites were grouped into regions representing the major zones of adaptation to drought according to CIMMYT's classification of mega-environments. Results spanning 1992 to 1997 were summarized on the basis of the number of times a particular site or region clustered with the target region, which was expressed as a fraction or percentage of the total number of possible groupings. Results indicated that the Centro de Investigaciones Agricolas del Noroeste (CIANO), CIMMYT's primary drought evaluation location, clustered with locations in South Asia, specifically India and Bangladesh. However, the number of clusters between CIANO and other Mexican locations with West Asia, Africa, and South America were fewer. This result suggests that the residual moisture stress generated at CIANO under limited irrigation conditions, while relevant to equivalent sites in the Indian Subcontinent, does not predict performance at locations where different stress patterns predominate. Associations among sites and regions, determined on the basis of clustering, ranged from weak (7% of total possible groupings in the case of Mexico and the Southern Cone of South America) to relatively strong (60% for Mexico and Bangladesh). Clusters of sites repeated in more than one year indicated two dominant groups, one for South Asian locations (including CIANO, Mexico) and another containing primarily South American sites.

DETERMINING the relationship among diverse yield testing environments and their degree of association is valuable in helping plant breeders better target germplasm to regions of broad or specific adaptation. The wheat (*Triticum aestivum* L.) breeding program of the International Maize and Wheat Improvement Center (CIMMYT) has developed and deployed the Semi-Arid Wheat Yield Trial (SAWYT) in many different environments around the world since 1992. This nursery is comprised of advanced bread wheat lines bred for tolerance to moisture stress.

CIMMYT's key drought evaluation site is located at the Centro de Investigaciones Agricolas del Noroeste (CIANO) in northwestern Mexico (27°20′N and elevation 38 m above sea level). Understanding the relationship between CIANO and key dry locations around the world is critical if we are to properly assess the effectiveness of this type of selection and evaluation. It is also important, particularly for CIMMYT's regional cooperators, to link the performance of different dry locations and regions from around the world with their own environments. Other authors have stated the importance of targeting germplasm to specific environments (Peterson and Pfeiffer, 1989) and increasing the efficiency of yield evaluation through the identification of key locations (Abdalla et al., 1996). Regions with similar dominant moisture stress patterns are the Southern Cone of South America, North Africa–West Asia–Southern Africa, and dry areas in South Asia (Rajaram et al., 1994; Calhoun et al., 1994).

Two types of multiplicative models have been used for studying genotype × environment interaction (GEI) and for developing methods for clustering sites or cultivars into groups without crossover interaction (COI) (Cornelius et al., 1992, 1993; Crossa et al., 1993, 1995, 1996; Crossa and Cornelius, 1993, 1997; Osman et al., 1997). These are the shifted multiplicative model (SHMM) in which $\bar{y}_{ij.} = \beta + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij.}$ (Seyedsadr and Cornelius, 1992) and the site regression model (SREG) in which $y_{ij.} = \mu_j + \sum_{k}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij.}$ (Cornelius et al., 1996). The variable $\bar{y}_{ij.}$ is the mean of the i^{th} cultivar ($i = 1,2, \ldots, g$) in the j^{th} environment ($j = 1,2, \ldots, e$); β is the shift parameter; μ_j is the site mean; λ_k ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$) are singular values that allow the imposition of orthonormality constraints on the singular vectors for cultivars, $\alpha_{ik} = (\alpha_{1k}, \ldots \alpha_{gk})$ and sites, $\gamma_{jk} = (\gamma_{1k} \ldots, \gamma_{ek})$, such that $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$ and $\sum_i \alpha_{ik}\alpha_{ik'} = \sum_j \gamma_{jk}\gamma_{jk'} = 0$ for $k \neq k'$; $\bar{\varepsilon}_{ij.}$ is the residual error.

If SHMM and SREG models with one multiplicative component ($SHMM_1$ and $SREG_1$) are adequate for fitting the data and primary effects of the sites, $\hat{\gamma}_{jl}$, all of like sign, then $SHMM_1$ and $SREG_1$ predict non-COI. Thus all cultivars should have consistent patterns of response across all sites included in the analysis (Crossa and Cornelius, 1997). On the contrary, if $\hat{\gamma}_{jl}$ are of different signs, then $SHMM_1$ and $SREG_1$ models predict COI, that is, cultivar ranking in the sites with negative $\hat{\gamma}_{jl}$ are the reverse of the cultivar ranking in the sites with positive $\hat{\gamma}_{jl}$.

This analysis has been used to determine environmental subgroups of large numbers of sites sown to the same set of cultivars (Fox et al., 1985, 1990). However, trials conducted over many years frequently contain unbalanced sets of cultivars as breeders constantly replace lines with newer materials. In this instance pattern analysis, a combination of classification and ordination analyses has been successfully employed (DeLacy and Lawrence, 1988; Peterson and Pfeiffer, 1989; Abdalla et al., 1996). These techniques have been used to examine the association of locations to CIMMYT spring bread wheat germplasm (DeLacy et al., 1994). However, all these

Wheat Program, International Maize and Wheat Improvement Center (CIMMYT) Apdo. Postal 6-641, 06600 Mexico DF, Mexico; J. Crossa, Biometrics and Statistics Unit, CIMMYT. Received 7 Sept. 2000. *Corresponding author (r.trethowan@cgiar.org).

Published in Crop Sci. 41:1461–1469 (2001).

Abbreviations: CIMMYT, International Maize and Wheat Improvement Center; SAWYT, Semi Arid Wheat Yield Trial; CIANO, Centro de Investigaciones Agricolas del Noroeste; GEI, genotype × environment interaction; COI, crossover interaction; SHMM, shifted multiplicative model; SREG, site regression model; SED, squared Euclidean distances.

Table 1. Locations returning yield data for the 1st through 6th SAWYTs in each geographic region.

Region	Site and country	SAWYT nursery number	Latitude	Altitude	Yield mean†	Standard deviation‡	SAWYTs reporting significant disease§
				m	Mg ha^{-1}		
Southern Africa							
1	Small Grain Inst. (S. Africa)	1,5,6	28°12′S	1687	2.42–5.02	0.52–0.66	1,5
2	Langgewens (S. Africa)	3	33°17′S	91	4.49	0.73	
3	Moredou (S. Africa)	2	33°18′S	82	2.88	0.68	
4	Gwebi (Zimbabwe)	2	17°41′S	1448	4.88	0.59	
North Africa							
5	Gezira (Sudan)	1	14°24′N	411	2.98	0.66	
6	El Khroub (Algeria)	1	36°7′N	640	3.56	0.42	
7	Sidi-Bel-Abbes (Algeria)	3	35°17′N	483	4.29	0.81	
8	Zakaria (Algeria)	1	34°36′N	980	0.60	0.17	
9	Khemis-Miliana (Algeria)	2	36°15′N	289	3.32	0.48	
10	Ain El Hadjar (Algeria)	2			1.37	0.34	
11	Beni-Slimane (Algeria)	3	36°14′N	600	2.59	0.43	
12	Ismailia (Egypt)	1,3	30°36′N	10	2.89, 3.72	0.77–0.84	3
13	El Qasser (Egypt)	2	31°20′N	2	1.31	0.41	
14	Boulifa (Tunisia)	1	36°38′N	350	4.12	0.31	
East Africa							
15	Kisozi (Burundi)	5	3°33′S	2090	1.34	0.37	
16	NPBRC–Njoro (Kenya)	5	0°25′S	2165	4.18	0.61	5
17	Bembeke (Malawi)	6	14°10′S	1560	0.48	0.10	
18	Holetta (Ethiopia)	6	9°3′N	2400	2.94	0.85	
19	Simba (Tanzania)	6	3°13′S	1750	2.78	0.40	
20	Lyamungo (Tanzania)	3	3°14′S	1280	2.08	0.33	
West Asia							
21	SPII Cereal Res. Inst. (Iran)	2	35°50′N	1321	5.56	0.75	
22	Zahak-zabol (Iran)	2	30°53′N	493	4.84	0.87	
23	Ahwaz (Iran)	2	31°17′N	20	5.07	0.89	
24	Gonbad (Iran)	2	37°16′N	76	4.19	0.71	
25	Ramtha (Jordan)	1	32°34′N	520	1.61	0.24	
26	Rawdat Harma (Qatar)	1	25°48′N	50	4.13	0.95	
27	ICARDA Tel-Hadya (Syria)	1,5	36°1′N	282	2.26, 4.19	0.57, 0.61	1
28	Shesham Bagd (Afghanistan)	2	34°25′N	552	4.68	0.44	
29	Al Kharj (Saudi Arabia)	2	24°15′N	540	3.88	0.58	
Central Asia							
31	Akmola¶ (Kazakstan)	4,5	51°10′N	300	1.34, 1.04	0.21, 0.33	
South Asia							
32	Dinajpur W.R.C (Bangladesh)	1,5	25°38′N	30	4.64, 3.29	0.88, 0.47	1
33	Rajshahi (Bangladesh)	2,3	24°22′N	18	3.42, 3.36	0.28, 0.51	
34	Bhairahawa (Nepal)	1,2,3,4,5	27°6′N	105	1.39–2.25	0.42–0.35	
35	Islamabad (Pakistan)	1,4	33°45′N	683	4.41, 3.62	*0.47, 0.60*	
36	Sariab (Pakistan)	2,5	30°12′N	1600	0.94, 1.52	0.16, 0.32	
37	Pirsabak (Pakistan)	5	33°59′N	340	3.36	0.56	
38	Wheat Res. Inst. (Pakistan)	4	31°25′N	117	3.00	0.87	
39	Barani (Pakistan)	4,5	32°56′N	490	2.36	0.25	
40	Dera Ismail Khan (Pakistan)	2,3,4,5,6	31°50′N	290	0.52–2.75	0.24–0.52	
41	Durgapura (India)	3	26°58′N	450	2.00	0.43	
42	DWR-Karnal (India)	3	29°40′N	300	5.65	0.69	
43	Vijapur (India)	3	23°35′N	126	2.59	0.52	
44	PAU–Ludhiana (India)	5	30°56′N	247	4.68	0.66	
East Asia							
45	San-Pa-Tong (Thailand)	3,4	18°30′N	300	0.88, 0.97	0.14, 0.34	
46	Suwan Farm (Thailand)	3	14°40′N	300	3.84	0.57	
47	Samoeng (Thailand)	4	18°17′N	820	3.45	0.86	
48	Pang Ma Pha (Thailand)	4	19°28′N	560	4.88	0.49	
North America							
49	CIANO (Mexico)	1,3,4,5,6	27°20′N	38	2.60–4.14	0.41–0.23	
50	El Batan (Mexico)	5	19°31′N	2249	3.65	0.64	
51	Mixteca Oaxaca (Mexico)	3,4	17°33′N	2250	2.71, 1.80	0.51, 0.44	
52	Tecamac (Mexico)	6	19°43′N	2260	2.64	0.82	
53	Tlacaque (Mexico)	6	19°45′N	2300	3.66	0.52	
54	San Franc. Atizapan (Mexico)	5	19°16′N	2640	3.41	0.45	
55	Kernen Res. Farm (Canada)	2,3,5,6	52°9′N	497	4.24–5.93	0.41–0.63	
56	Swift Current (Canada)	2	50°17′N	825	3.16	0.25	
Southern Cone							
57	Bela Vista (Brazil)	1	23°0′S	618	2.07	0.38	
58	Londrina (Brazil)	2	23°22′S	540	3.21	0.41	
59	CNP-Soja (Brazil)	2	23°12′S	620	4.06	0.56	
60	Pergamino (Argentina)	1,3,6	33°56′S	65	2.77–5.01	0.52–0.23	3
61	Marcos Juarez (Argentina)	1,3,4,5,6	32°42′S	110	1.19–2.79	0.34–0.38	
62	Tucuman-Obispo (Argentina)	1,6	26°48′S	460	1.23, 2.47	0.21, 0.42	
63	Bordenave (Argentina)	1	37°51′S	212	5.12	0.99	
64	Parana (Argentina)	1	31°50′S	110	2.03	0.12	1
65	Cordoba (Argentina)	2	31°30′S	425	0.38	0.15	
66	La Tijereta (Argentina)	5	32°8′S	200	2.26	0.56	5

Continued next page.

Table 1. Continued.

Region	Site and country	SAWYT nursery number	Latitude	Altitude	Yield mean†	Standard deviation‡	SAWYTs reporting significant disease§
Andean Region							
67	San Benito (Bolivia)	1,2,3,4,5,6	17°30′S	2730	0.88–3.58	0.14–0.72	4
68	Sta. Catalina (Ecuador)	1,5	0°22′S	3050	3.35, 2.95	0.44, 0.33	1
Central America							
69	Zamorano (Honduras)	4	14°0′N	805	1.16	0.27	
Southern Europe							
70	Gimenells (Spain)	5	41°35′N	290	4.16	0.67	
71	La Mojonera (Spain)	1	39°58′N	220	5.28	0.49	
72	Cent. de Inv. Agrario (Spain)	2	37°21′N	650	4.35	0.43	
73	Torregrossa/Belloc (Spain)	2,4	41°35′N	200	4.88, 3.41	0.69, 0.75	
74	Cortijo Torrevuelo (Spain)	2	37°38′N	72	0.96	0.14	
75	La Orden (Spain)	4	38°49′N	200	3.38	0.69	
76	Tobiscal (Spain)	4			4.19	0.59	
78	Kentziko Thermi (Greece)	4	40°38′N	38	3.52	0.46	
Eastern Europe							
79	Szeged (Hungary)	1	46°0′N	80	4.30	0.48	
80	Odessa (Ukraine)	3	46°27′N	34	3.48	0.83	
81	Kharkiv (Ukraine)	4	50°0′N	170	1.15	0.31	
82	Spring Wheat Lab. (Russia)	3,4	53°1′N	47	1.54, 2.70	0.37, 0.43	
Total sites		122					

† When more than two SAWYTs are sown at the same site in different years a yield range is presented.
‡ When more than two SAWYT are sown at the same site the standard deviations of the lowest and highest yielding sites are presented.
§ Diseases reported included stem, leaf and stripe rust and *Septoria tritici*.
¶ Akmola has recently been renamed Astana.

cultivars were developed for irrigated conditions, and site associations were determined across both irrigated and low rainfall conditions. There has been no such attempt to classify global drought locations sown to cultivars specifically developed for performance under moisture limiting conditions.

The aim of this paper is to (i) examine the relevance of selection under terminal moisture stress at CIANO, Mexico compared to the primary drought affected target areas around the world and (ii) examine the association among international testing locations where the SAWYT nursery is planted.

MATERIALS AND METHODS

Locations and Cultivars

Yield data from a total of 156 locations were returned from the SAWYT between 1992 and 1997. A total of six yield nurseries (SAWYTs 1–6), each comprised of 30 to 50 cultivars, were sown. Although most cultivars varied from year to year, a local check cultivar representing the best locally adapted germplasm was included at each site each year. The local check cultivar varied among locations and in some instances changed between years at the same location. All trials were sown as two replicate alpha-lattice designs (Barreto et al., 1997). Yield data from each trial were analyzed by SAS (SAS, 1988). Genotypes were considered fixed effects and replicates and subblocks within replicates as random effects. Adjusted means were calculated for subsequent SHMM and pattern analyzes. Trials were sown under a range of different moisture conditions. A site is defined as a location/year occurrence. High yielding irrigated locations (arbitrarily defined as those with means above 6 Mg ha^{-1}) were removed to ensure that the remaining sites were representative of the potential yield range in most water limited locations. To ensure clusters among locations were biologically based and not artefactual, only those sites indicating significant differences among genotypes, regardless of the size of the coefficient of variation, were retained giving a total number of 122 sites (Table 1). Diseases scored were stem, leaf, and stripe rust (caused by *Puccinia striiformis* f. sp. *hordeii*) and *Septoria tritici* Roberge in Desmaz.

The sites were grouped into seven regions representing northern, southern and eastern Africa, West Asia, South Asia, the Southern Cone of South America, and Mexico. These groupings represent regions suffering somewhat different stress patterns as determined on the basis of long-term weather records. North Africa, West Asia, and Southern Africa generally experience Mediterranean or postanthesis moisture stress; South Asia experiences residual moisture stress; and the Southern Cone predominantly preanthesis stress (Rajaram et al., 1994). Rainfall records were incomplete for a majority of locations during the years in which the SAWYT was grown, for this reason we relied on long term regional rainfall averages to determine into which broad category a particular site fell. To provide separate comparisons of all other regions with the Mexican evaluation sites, the latter were classified as a distinct region. Many locations, particularly in West Asia and North Africa, were sown to the SAWYT trial only once during the 6-yr period. For this reason individual trials within these regions were considered collectively in comparisons with other regions and locations.

Germplasm entering the SAWYT was developed in Mexico by shuttling segregating materials between two contrasting moisture regimes (Rajaram et al., 1994). At CIANO severe terminal moisture stress was generated during the winter crop cycle by gravity irrigating preformed beds 14 d prior to sowing. Segregating and advanced lines were sown in November on a receding moisture profile with no subsequent irrigation. Twenty-year average annual rainfall for the cropping period November to April is 48.2 mm. Materials were harvested in April and sown in May at Toluca in the central Mexican highlands (19°16′N and 2640 m above sea level) which receive approximately 800 mm of annual precipitation. Under this environment, materials are selected for responsiveness to moisture, nutrient inputs, and resistance to disease.

Analysis and Grouping of Locations

Multiplicative Models for Clustering Sites Without Crossover Interaction

In various site-clustering procedures developed on the basis of SHMM or SREG (Cornelius et al., 1992; Crossa et al., 1993; Crossa and Cornelius, 1997), the measure of distance (i.e., dis-

similarity) between a pair of sites is the residual sum of squares (RSS) after fitting $SHMM_1$ or $SREG_1$, RSS($SHMM_1$) or RSS ($SREG_1$), respectively. The dichotomous splitting procedure used on the dendrogram obtained from SHMM cluster analysis facilitates finding groups with negligible COI within clusters. Computations are facilitated because the site regression model with one multiplicative term can be reparameterized as a shifted multiplicative model with one multiplicative component. In this study, the SHMM clustering procedure for grouping sites without COI (Crossa et al., 1993) was applied to each of the six SAWYTs, and clusters of sites with negligible COI were found.

Pattern Analysis

Pattern analysis is the clustering and ordination of sites (or/and cultivars) in the two-way data table of cultivars × sites. In this study, the data used were the three-way table of cultivar × site × year. It was assumed that cultivars in any given year were a representative sample of the germplasm under evalua-

Table 2. Summary of regional associations from dendrograms generated for each SAWYT.

Number of groupings with the region/Total number of possible groupings

Mexican Region								
Sites†	CIANO	Semillas	Mixteca	Atizapan	El Batan	Temacac	Rancho	Total
Bangladesh	4/6	0/1	1/1	1/1	0/1	–	–	6/10
India	2/4	–	2/3	0/1	1/1	–	–	5/9
South Asia	9/25	0/3	4/11	2/7	1/7	0/1	0/1	16/55
Brazil	1/4	0/1	–	–	–	0/1	1/1	2/7
South Africa	1/3	0/1	1/1	0/1	0/1	–	–	2/7
Nepal	1/4	0/1	0/2	1/1	0/1	–	–	2/9
Mexico	2/7	0/1	1/2	1/2	0/2	0/2	0/2	4/18
N. Africa	1/8	2/3	0/3	–	–	–	–	3/14
Bolivia	0/5	1/1	0/2	0/1	0/1	0/1	1/1	2/12
Argentina	3/24	2/5	0/6	0/2	0/2	0/3	2/3	7/45
Spain	1/6	0/2	1/3	0/1	0/1	–	–	2/13
Pakistan	2/11	0/1	1/5	0/4	0/4	0/1	0/1	3/27
West Asia	1/4	0/3	–	0/1	0/1	–	–	1/9
Total	28/111	5/23	11/37	5/22	2/22	0/9	3/9	

South Asian Region					
Sites	Banglad.	Pakistan	Nepal	India	Total
Kazakstan	0/2	7/12	1/3	0/2	8/19
South Africa	0/2	5/9	0/4	2/4	7/19
Bangladesh	–	0/9	3/4	2/4	5/17
Mexico	4/6	3/23	1/8	5/8	13/45
West Asia	4/10	3/28	4/11	0/1	13/50
India	2/4	1/7	0/4	2/6	5/21
Argentina	1/10	9/25	1/11	2/8	13/54
Canada	0/3	4/9	0/3	0/4	4/19
South Asia	5/17	7/56	3/21	5/21	20/115
N. Africa	2/11	2/17	4/11	0/9	8/48
Bolivia	0/4	3/14	0/5	1/4	4/27
Spain	1/5	4/24	1/9	0/1	6/39
Nepal	3/4	0/13	–	0/4	3/21
Pakistan	0/9	6/27	0/13	1/7	7/56
Total	22/87	54/273	18/107	20/83	

West Asian Region							
Sites	Iran	Jordan	Qatar	Syria	Afghani	Saudi A.	Total
Bangladesh	0/4	1/1	0/1	1/2	1/1	1/1	4/10
Nepal	0/4	1/1	0/1	1/2	0/1	0/1	2/10
South Asia	2/20	2/3	0/3	3/10	1/5	1/5	9/46
West Asia	0/3	1/2	0/2	1/2	1/6	1/6	4/21
N. Africa	1/12	1/5	1/5	1/5	1/3	1/3	6/33
Algeria	1/8	0/2	0/2	0/2	1/2	1/2	3/18
Mexico	–	0/2	1/2	0/4	–	–	1/8
Pakistan	2/12	0/1	0/1	1/5	0/3	0/3	3/25
South Africa	1/4	0/1	0/1	0/2	0/1	0/1	1/10
Argentina	0/4	0/5	0/5	0/7	1/1	1/1	2/23
Spain	0/12	0/2	0/2	0/3	1/3	1/3	2/25
Brazil	0/8	0/1	1/1	0/1	0/2	0/2	1/15
Total	7/91	6/26	3/26	8/45	7/28	7/28	

North African Region					
Sites	Sudan	Algeria	Egypt	Tunisia	Total
Nepal	0/1	1/6	1/3	1/1	3/11
Spain	0/2	3/10	2/5	0/2	5/19
Argentina	0/5	5/16	2/8	0/5	7/34
Bangladesh	0/1	1/6	0/3	1/1	2/11
Bolivia	0/1	2/6	0/3	0/1	2/11
Spain	0/2	2/10	2/8	0/2	4/22
Brazil	1/1	0/6	1/3	0/1	2/11
West Asia	1/3	3/20	0/17	2/3	6/43
South Asia	0/3	2/28	3/18	2/3	7/52
N. Africa	0/4	3/16	1/10	0/4	4/34
Pakistan	0/1	0/10	2/9	0/1	2/21
South Africa	0/1	0/6	1/3	0/1	1/11
Mexico	1/2	0/8	0/4	0/2	1/16
Total	3/27	22/148	15/94	6/27	

Continued next page.

tion. Sites (individual location/year occurrences) were judged on the basis of their ability to discriminate among cultivars. Only sites that occurred in two or more SAWYTs were included in the overall pattern analysis. Since some sites were sown to more than two SAWYTs in different years, their comparisons had different levels of precision. The clustering strategy used is that recommended by (DeLacy and Cooper, 1990) and used by Abdalla et al. (1996). Dissimilarities between sites in each year and averaged across years were measured by squared Euclidean distances (SED). Since different years had different numbers of cultivars, the averages were weighted by the number of cultivars in each year. The incremental sum of squares criterion and the agglomerative hierarchical strategy procedure with SED as the dissimilarity measure were used for classification.

RESULTS AND DISCUSSION

Associations among all Locations and Regions

Average site yields ranged from 0.38 to 8.48 Mg ha^{-1} during the 6-yr period. Significant disease incidence was reported at 11 of the 122 sites included in the analysis and no sites reported insect damage. Dendrograms developed from the SHMM cluster analysis were used to examine the association of various sites with key regions that frequently experience drought and with sites within those regions. A summary of dendrogram results is presented in Table 2. The number of clusters among various sites located within the seven key regions is expressed as a fraction of the total number of possible groupings or clusters. Site clusters were determined at the third fusion or third group level of the SHMM cluster analysis. A total value for the region is also calculated and expressed as a fraction. For example, comparisons of India with the Mexican region had a total value of 5/9 (56%). This was calculated by adding the fractions of the individual groups CIANO (2/4), Mixteca (2/3), Atizapan (0/1) and El Batan (1/1). Similarly, each location within each of the seven key regions is totaled across locations that clustered at least once with locations in each key region. For example, CIANO, which appears in the Mexican region, clustered 28 times out of 111 possible groupings with 13 different global locations or regions.

The South Asian Region

Rainfall, soil type, and farming practice in this region are diverse. Nepalese sites cluster 14% (three of 21 possible groupings) of the time with other sites across

Table 2. Continued.

Number of groupings with the region/Total number of possible groupings

Southern African Region

Sites	S. Africa	Zimbab.	Total
Kazakstan	2/2	–	2/2
Canada	3/4	–	3/4
Argentina	6/13	0/1	7/14
India	2/4	–	2/4
Pakistan	5/10	0/3	5/13
South Asia	8/22	0/5	8/27
Mexico	2/9	–	2/9
Bangladesh	1/4	0/1	1/5
Bolivia	0/5	1/1	1/6
Spain	1/6	0/3	1/9
West Asia	1/11	0/7	1/18
Total	31/90	1/21	

Eastern African Region

Sites	Tanzania	Burundi	Kenya	Malawi	Ethiopia	Total
West Asia	–	1/1	1/1	–	–	2/2
N. Africa	2/3	–	–	–	–	2/3
E. Africa	0/2	1/1	1/1	0/2	0/2	3/8
Pakistan	1/2	1/4	1/4	0/1	0/1	4/12
Nepal	1/1	0/1	0/1	–	–	1/3
Brazil	0/1	–	–	0/1	0/1	1/3
Argentina	1/5	0/2	0/2	0/3	0/3	4/15
South Asia	2/8	1/7	1/7	0/1	0/1	5/24
Mexico	1/5	0/2	0/2	1/3	0/3	3/15
Bolivia	0/2	0/1	0/1	0/1	0/1	1/6
Total	8/29	4/19	4/19	1/12	0/12	

Southern Cone of South America

Sites	Brazil	Argentin	Total
Kazakstan	–	2/3	2/3
Canada	1/3	3/8	4/11
Argentina	3/10	8/26	11/36
Algeria	1/6	5/14	6/20
Bolivia	0/4	5/13	5/17
Pakistan	0/8	9/24	9/32
Spain	2/6	5/21	7/27
N. Africa	3/11	7/29	10/40
South Africa	0/4	3/9	3/13
Brazil	–	2/9	2/9
South Asia	2/15	10/53	12/68
Bangladesh	1/3	1/9	2/12
West Asia	3/17	2/20	5/37
Mexico	1/5	1/25	2/30
Nepal	1/4	0/13	1/17
Total	18/96	63/276	

† Sites include countries and geographic regions.

the South Asian Region. However, within this region, Nepal and Bangladesh are closely associated, while no relationship exists between Nepal and Pakistan. Kazakstan, South Africa, Canada, and Argentina also associate with Pakistan but less so with other countries within the South Asian region. West Asian locations cluster better with Bangladesh and Nepal than with Pakistan. Mexican locations, including CIANO, are the best predictors of environments in India and Bangladesh; however, they associate poorly with most sites in Pakistan and Nepal (Table 2). Even though the variation in latitude within the region is not large (24–33°N), stress patterns across the region are very different. In Pakistan, dry areas are less influenced by monsoonal rain than are equivalent areas in central India, while in Nepal and some parts of India and Bangladesh stress patterns are influenced by availability of irrigation. Stress patterns in Pakistan are similar to those in higher latitude areas such as Kazakstan and Canada (Table 2). This association reflects the lack of photoperiod response in the CIMMYT materials included in the SAWYT nurseries. The grouping of South Africa and Pakistan, both dry rainfed areas of equivalent latitude, indicates a significant degree of association between these areas. Nepal and Bangladesh cluster together because of similar limited irrigation production systems. Within the South Asian region, Bangladesh and India clustered best with 14 different global sites and regions (Table 2).

The West Asian Region

Four Iranian sites returned yield data for the 2nd SAWYT, making it the best represented country in the West Asian region. Unfortunately, the 2nd SAWYT was not sown in Mexico, so comparisons between these sites and Mexican locations cannot be made. Iranian locations cluster least with other sites in the West Asian region and show very little association with other global locations and regions (Table 2). The two locations in Pakistan giving the closest, although still weak association with Iran, Sariab, and Barani are located in the dry, northern areas. If Iran is eliminated, then Bangladesh becomes the best predictor of West Asian locations (4/6), clustering with Jordan, Syria, Afghanistan, and Saudi Arabia followed by the combined South Asian region. Those sites clustering with Bangladesh are rainfed locations, with the exception of Jordan and Saudi Arabia, where the trial was sown under limited irrigation. The limited irrigation regimes generated in Bangladesh, therefore, appear to mimic those in the terminal moisture stress environments of West Asia. Lack of reliable rainfall records from these West Asian locations in the years the trials were sown make it difficult to draw firm conclusions. With removal of the Iranian sites, the next best predictors of West Asia are West Asian locations themselves (4/18) and North African sites (5/21), all of which have, based on long-term rainfall averages, similar Mediterranean type stress patterns. Mexican and South American locations clustered poorly with West Asia, ranging from 1/15 in Brazil to 1/8 in Mexico. Afghanistan and Saudi Arabian locations within the West Asian region clustered best across global locations (Table 2).

The North African Region

Clusters of North African sites with global sites and regions indicated that Nepal gave the best association, followed by Spain and the South American sites in Argentina, Bolivia, and Brazil (Table 2). Mexican sites did not predict this region well (1/16). The Sudanese site expressed little if any association with any of the SAWYT sites. This may be due severe heat stress late in the growing cycle, which is often experienced in Sudan and other North African locations. The best associations within the region are between Algeria and Argentina and Algeria and Bolivia (Table 2). These sites experience early season, preanthesis drought stress. Within the North African region Tunisia clustered best with other global sites and regions.

The Southern African Region

This region is comprised of the two countries South Africa and Zimbabwe. The association of sites in this region may be inflated because only six locations were used. Zimbabwe showed a different pattern of association compared with the South African sites (Table 2). This reflects the different latitude of the Zimbabwean site (17°S) compared with the South African locations (28–33°S). The grouping of high latitude locations in Kazakstan and Canada with South Africa reflects similar stress patterns. The grouping of Pakistan, India and Argentina suggested that stress patterns were similar among these regions. It is expected that West Asian and North African sites, which experience Mediterranean type drought stress, would group with South Africa. However, only one out of 22 possible groupings occurred between southern Africa and regions with similar stress patterns. However, the strong association of sites in South Africa with 11 different locations (31/90) from around the world suggests that South Africa could be utilized in global wheat breeding efforts.

The Eastern African Region

In this region, four locations returned data from one year only and a fifth, Tanzania, reported data from two years. Stress patterns in Eastern Africa tend to be similar to those in West Asia and North Africa (4/5), indicating the presence of terminal or late season drought stress. Tanzania was the only eastern African location to cluster with CIANO (1/6, not shown in Table 2). South Asian locations where farmers plant on residual moisture following the monsoon or use limited irrigation and Southern Cone locations, typified by preanthesis stress did not associate well with eastern Africa. Ethiopia (0/12) and Malawi (1/12) did not associate well with other global locations and regions. Rainfall records are not available for the Eastern African sites in the years the trials were sown making it difficult to assess whether the growing conditions were different from the long-term average. Among the eastern African locations,

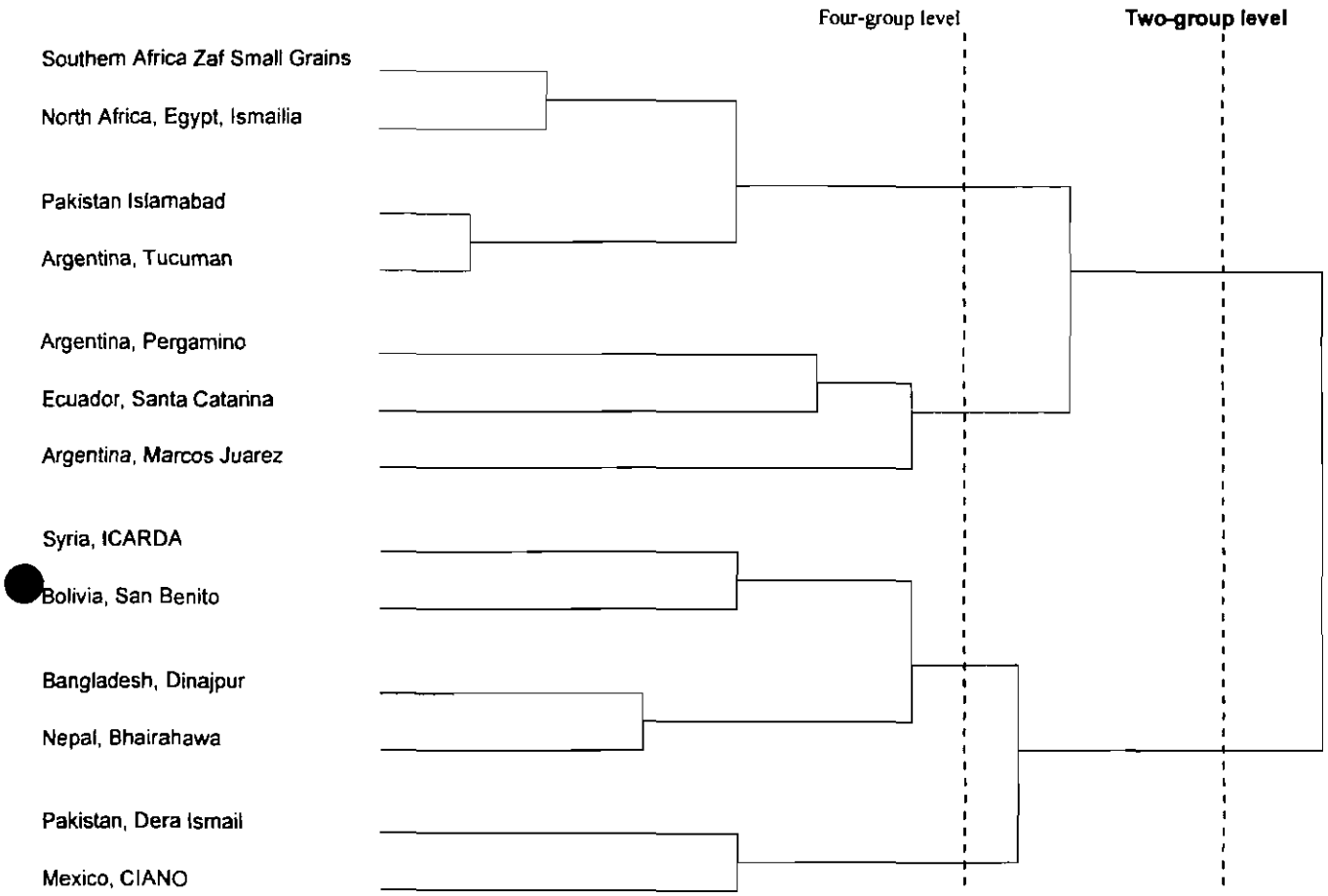

Fig. 1. Dendrogram of the relationships among sites sown to the SAWYT in more than 1 yr.

Tanzania was most closely associated with other global locations followed by Burundi and Kenya (Table 2).

The Southern Cone of South America

This region is represented by locations in Argentina and Brazil. Their association with high latitude sites in Kazakstan and Canada would indicate similar patterns of adaptation (Table 2). The grouping of Pakistan, Bolivia, Algeria, and Spain with Argentina indicates that patterns of adaptation in Argentina are similar to many parts of the world, where sites experience preanthesis drought stress. Limited rainfall records were available for some sites in Argentina during the years covering this study. While these records indicate preanthesis drought stress predominated, significant rainfall variation was observed at many sites; this is reflected in the relatively weak grouping of locations in Argentina with each other (8/26). Other South Asian sites Nepal, India, and Bangladesh most of which are sown on residual moisture or under limited irrigation, did not associate with Argentina (1/29). Mexico, where development of the genotypes and most of the yield trials were sown using a single preseeding irrigation showed very little association with this region (2/30). The clustering of Argentinean locations with many different global sites and regions (63/276) highlights the strategic value of these locations in differentiating germplasm in a global breeding program.

Association of Locations Repeated in More than One Year

Pattern analysis was used to examine the association among sites repeated in more than one year (Fig. 1). At the first fusion level two groupings resulted. Group 1 contained three locations from Argentina and one from Santa Catalina in Ecuador. South Africa, Egypt, and one location in Pakistan made up the remaining sites. Group 2 contained three South Asian locations, Bangladesh, Nepal, and Pakistan. In addition, Syria, Bolivia, and CIANO (Mexico) also clustered in this group.

Pattern analysis confirmed the conclusions of individual SHMM analyses by indicating an association between CIANO and South Asia (Nepal, Pakistan, and Bangladesh). As no Indian site was sown to SAWYT for more than one year, we could not include India in the comparison. Most South American locations, with the exception of Bolivia, clustered together in a separate group, also supporting SHMM conclusions. Interestingly, Syria, the only West Asian location reporting data for more than one year, also grouped with CIANO, indicating a degree of similarity between stress patterns at these two sites.

Association of Global Locations with CIANO

Locations grouping with CIANO on the basis of yield at the third fusion level of the SHMM cluster analysis

Table 3. Sites grouping at the third fusion level in the SHMM cluster analysis with CIANO, Mexico, for yield from the 1st, 3rd, 4th, 5th, and 6th SAWYTs.

Sites clustering with CIANO	No. of times grouping with CIANO
1st SAWYT	
Sudan; Gezira Res. Station	1/1
Brazil; Bela Vista Do Paraiso	1/1
Qatar; Rawdat Harma	1/1
3rd SAWYT	
India; Durgapura	1/1
India; Vijapur	1/1
Mexico; Mixteca Oaxaca	
Bangladesh; Rajshahi	1/1
4th SAWYT	
Portugal; PBS Alentejo	1/1
Ukraine; Odessa	1/1
Pakistan; NARC Islamabad	1/2
5th SAWYT	
Mexico; San Franc. Atizapan	1/1
Bangladesh; Dinajpur	1/2
Nepal; NWRP, Bhairahawa	1/4
6th SAWYT	
Tanzania; Simba	1/1
Pakistan; Dera Ismail Khan	1/4
Argentina; Tucuman-Obispo	1/2

were determined. Many of these locations were sown only once to the SAWYT and are indicated in Table 3. Sites in Sudan, Brazil, Qatar, India, Bangladesh, Portugal, Ukraine, Mexico (Atizapan), and Tanzania reported SAWYT data only once and clustered with CIANO. Other sites in Mexico (Oaxaca), Pakistan (NARC), Bangladesh (Dinajpur), and Argentina appeared twice, clustering once with CIANO, while two locations in Nepal and Pakistan (Dera Ismail Khan) occurred four times, clustering only once with the CIANO site.

Because many locations clustering with CIANO from the SHMM analyses were sown only once to the SAWYT during the 6-yr period, we can only conclude that an association exists in the year in which the head-to-head comparison occurred. The repeatability of these associations in many cases cannot be determined. The grouping of Indian and Bangladesh sites with CIANO could be explained by similarities in latitude (Table 1) and moisture stress patterns. The drought stress generated at CIANO under limited irrigation is similar to that experienced in many parts of South Asia.

At CIANO, late season severe terminal drought stress is generated by the application of a single preseeding irrigation. This screening method is designed to mimic the terminal moisture stress experienced in the South Asian region following the monsoon. In these regions farmers plant after the monsoonal rains on a receding moisture profile: very little if any rain falls after sowing. The slightly higher latitude and altitude of most Pakistani sites (between 4 and 7° farther north and between 79 and 1562 m higher) may explain the reduced level of association between these areas and CIANO. The site most similar in latitude and altitude, Nepal, clustered only 25% of the time with CIANO, however, pattern analysis based on repeated sites outlined in the previous section does confirm a relationship between CIANO and Nepal.

Not surprisingly, only two other Mexican locations clustered with CIANO out of a total of seven possible comparisons. Apart from CIANO, all these sites are located at 2249 to 2640 m above sea level, compared with CIANO's altitude of 38 m, and are at least 8° latitude closer to the equator. Two higher latitude sites, Portugal (38°N) and Ukraine (46°N) also clustered with CIANO.

The low level of grouping between CIANO and Southern Cone locations in Brazil and Argentina (Table 2) can be explained by different prevailing moisture conditions. These sites experienced preanthesis drought stress throughout the duration of the study. Therefore it is not surprising that SAWYT genotypes, developed under moderate to severe terminal moisture stress, differentiated differently for yield in the Southern Cone. This is borne out by the much stronger association between CIANO and India and Bangladesh (Table 2). Sites in India and Bangladesh under limited irrigation generally apply all the available water prior to anthesis. However, the stress generated at CIANO did not associate well with West Asian and North African sites. A possible explanation is that many of the sites in these regions are generally cooler than CIANO and genotype ranking may be influenced by the longer growing season.

Association Between the Same Location Sown to the Same Genotypes in Different Years

A small number of genotypes, ranging from 5 to 10, were in common between years in comparisons between specific SAWYT trials (Table 4). Dendrograms (not

Table 4. The number of times CIANO and other locations sown to different SAWYTs in different years cluster with themselves at the third fusion level in the SHMM analysis based on genotypes common to each comparison.

	Number of groupings/Total number of possible groupings						
SAWYT comparison†	Mexico CIANO	Nepal Bhairahwa	Bolivia San Benito	Canada Kernen	Argentina M. Juarez	Argentina Pergamino	Pakistan Dera Ismail
1 v 2	–	0/1	0/1	–	–	–	–
2 v 3	–	0/1	0/1	1/1	–	–	1/1
3 v 4	1/1	0/1	0/1	–	0/1	–	0/1
4 v 5	0/1	1/1	1/1	–	1/1	–	1/1
5 v 6	1/1	–	0/1	1/1	0/1	–	0/1
1 v 3	1/1	1/1	0/1	–	0/1	0/1	–
3 v 5	1/1	1/1	0/1	1/1	0/1	–	0/1
3 v 6	0/1	–	0/1	0/1	1/1	0/1	0/1
4 v 6	1/1	–	0/1	–	0/1	–	0/1
Total	5/7	3/6	1/9	3/4	2/7	0/2	2/7

† Genotypes in common for each comparison ranged from 5 to 10.

shown) developed from SHMM cluster analysis indicated that CIANO, CIMMYT's primary drought testing location, clustered with itself on 5 of 7 occasions at the third fusion level. Sites in Nepal (3/6) and Canada (3/4) also indicated a relatively high degree of association between years. However, Bolivian, Argentinean, and Pakistani locations did not cluster to a significant degree with themselves in the paired comparisons among different SAWYTs.

The high degree of association between years of CIANO indicates this location has high repeatability in discriminating germplasm. However, the weak association between the Bolivian, Argentinean, and Pakistani locations reflects the inherent variability characteristic of most rainfed, drought prone environments. Disease was not a major factor influencing genotypic ranking and subsequent location clustering as only 11 sites out of 122 included in the analysis reported significant disease incidence (Table 1). The high degree of association observed between years for CIANO reflects the controlled irrigation conditions under which materials are grown.

CONCLUSIONS

As suggested by Osman et al. (1997), the SHMM analysis can be applied routinely to identify subsets of locations without COI and thus find key locations. Furthermore, results obtained by yearly SHMM analyses can always be confirmed by pattern analysis on long-term multilocation trial data. The primary aim of our study was to examine the relevance of selecting germplasm under controlled irrigation at CIANO, Mexico, compared with performance under global drought stressed environments. The relatively low degree of repeatability of these locations, heavily influenced by rainfall and other seasonal factors, make it difficult to find high levels of association between CIANO and many global locations year to year. However, the application of SHMM and pattern analyzes indicated that CIANO is similar to sites in India and Bangladesh. These are not typically rainfed areas and crops are frequently drought stressed through lack of irrigation water.

The gravity-fed residual moisture stress generated at CIANO may need to be modified to improve the relevance of materials selected in Mexico to locations in the Southern Cone, West Asia, and Africa. Generation of a combination of both terminal and preanthesis stress scenarios may increase the frequency of elite materials adapted to these regions. The secondary aim of our study was to examine relationships among international testing locations with a view to identifying key locations for drought screening. Selection using a combination of environments such as CIANO, South Africa, and Argentina, all of which correlated well across many different environments, may provide the platform for greater rates of progress in breeding for dry environments globally.

REFERENCES

Abdalla, O.S., J. Crossa, E. Autrique, and I.H. DeLacy. 1996. Relationships among International testing sites of spring durum wheat. Crop Sci. 36:33–40.

Barreto, H.J., G.O. Edmeades, S.C. Chapman, and J. Crossa. 1997. The alpha lattice design in plant breeding and agronomy: Generation and analysis. p. 544–551. *In* G.O. Edmeades et al. (ed.) Developing drought- and low N-tolerant maize. Mexico, D.F., Mexico.

Calhoun, D.S., G. Gebeyehu, A. Miranda, S. Rajaram, and M. van Ginkel. 1994. Choosing evaluation environments to increase wheat grain yield under drought. Crop Sci. 34:673–678.

Cornelius, P.L., J. Crossa, and M.S. Seyedsadr. 1996. Statistical tests and estimators for multiplicative models for cultivar trials. p. 199–234. *In* M.S. Kang et al. (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Cornelius, P.L., M.S. Seyedsadr, and J. Crossa. 1992. Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theor. Appl. Genet. 84:161–172.

Cornelius, P.L., D.A.van Sanford, and M.S. Seyedsadr. 1993. Clustering cultivars into groups without rank-change interactions. Crop Sci. 33:1193–1200.

Crossa, J., and P.L. Cornelius. 1993. Recent developments in multiplicative models for cultivar trials. p. 571–577. *In* D.R. Buxton et al. (ed.) International crop science I. CSSA, Madison, WI.

Crossa, J., and P.L. Cornelius. 1997. Site regression and shifted multiplicative models clustering of cultivar trials sites under heterogeneity of error variances. Crop Sci. 37:406–415.

Crossa, J., P.L. Cornelius, K. Sayre, and J.I. Ortiz-Monasterio. 1995. A shifted multiplicative model fusion method, for grouping environments without cultivar rank change. Crop Sci. 35:54–62.

Crossa, J., P.L. Cornelius, and M.S. Seyedsadr. 1996. Using the shifted multiplicative model cluster methods for crossover genotype-by-environment interaction. p. 175–198. *In* M.S. Kang et al. (ed.) Genotype-by-environment interaction. CRC Press, Boca Raton, FL.

Crossa, J., P.L. Cornelius, M. Seyedsadr, and P. Byrne. 1993. A shifted multiplicative model cluster analysis for grouping environments without genotypic rank change. Theor. Appl. Genet. 85:577–586.

DeLacy, I.H., and M. Cooper. 1990. Pattern analysis for the analysis of regional variety trials. p. 301–334. *In* M.S. Kang (ed.) Genotype-by-environment interaction in plant breeding. Louisiana State Univ., Baton Rouge, LA.

DeLacy, I.H., P.N. Fox, J.D. Corbett, J. Crossa, S. Rajaram, R.A. Fischer, and M. van Ginkel. 1994. Long-term association of locations for testing spring bread wheat. Euphytica 72:95–106.

DeLacy, I.H., and P. Lawrence. 1988. Combining pattern analysis over years—Classification of locations. p. 175–176. *In* K.S. McWhirter et al. (ed.) Proc. of the Ninth Australian Plant Breeding Conference. Agricultural Research Institute, Wagga Wagga, New South Wales, Australia.

Fox, P.N., A.A. Rosielle, and W.J.R. Boyd. 1985. The nature of genotype × environment interactions for wheat yield in Western Australia. Field Crops Res. 11:387–398.

Fox, P.N., B. Skovmand, B.K.Thompson, H.J. Braun, and R. Cormier. 1990. Yield and adaptation of hexaploid spring triticale. Euphytica 47:57–64.

Osman, S.A., J. Crossa, and P.L. Cornelius. 1997. Results and biological interpretation of shifted multiplicative model clustering of durum wheat cultivars and testing sites. Crop Sci. 37:88–97.

Peterson, C.J., and W.H. Pfeiffer. 1989. International winter wheat evaluation: Relationships among test sites based on cultivar performance. Crop Sci. 29:276–282.

Rajaram, S., M. van Ginkel, and R.A. Fischer. 1994. CIMMYT's wheat breeding mega-environments (ME). p. 1101–1106. *In* Proceedings of the 8th International Wheat Genetics Symposium. Beijing, China.

Seyedsadr, M., and P.L. Cornelius. 1992. Shifted multiplicative model for nonadditive two-way tables. Comm. Stat. B. Simulation and Computation 21:807–822.

SAS. 1988. SAS Institute Inc., Cary, NC.

Two Types of GGE Biplots for Analyzing Multi-Environment Trial Data

Weikai Yan,* Paul L. Cornelius, Jose Crossa, and L. A. Hunt

ABSTRACT

SA genotype main effect plus genotype × environment interaction (GGE) biplot graphically displays the genotypic main effect (G) and the genotype × environment interaction (GE) of the multienvironment trial (MET) data and facilitates visual evaluation of both the genotypes and the environments. This paper compares the merits of two types of GGE biplots in MET data analysis. The first type is constructed by the least squares solution of the sites regression model ($SREG_2$), with the first two principal components as the primary and secondary effects, respectively. The second type is constructed by Mandel's solution for sites regression as the primary effect and the first principal component extracted from the regression residual as the secondary effect ($SREG_{M+1}$). Results indicate that both the $SREG_2$ biplot and the $SREG_{M+1}$ biplot are equally effective in displaying the "which-won-where" pattern of the MET data, although the $SREG_2$ biplot explains slightly more GGE variation. The $SREG_{M+1}$ biplot is more desirable, however, in that it always explicitly indicates the average yield and stability of the genotypes and the discriminating ability and representativeness of the test environments.

MULTIENVIRONMENT TRIALS are conducted for all major crops throughout the world. The main purpose of MET is to identify superior cultivars for recommendation to farmers and to identify sites that best represent the target environment. Usually, a large number of genotypes are tested over a number of sites and years, and it is often difficult to determine the pattern of genotypic responses across environments without the help of graphical display of the data.

Yan et al. (2000) developed a "GGE biplot" methodology for graphical analysis of MET data. "GGE" refers to the genotype main effect (G) plus the genotype × environment interaction (GE), which are the two sources of variation that are relevant to cultivar evaluation. A biplot (Gabriel, 1971) is a plot that simultaneously displays both the genotypes and the environments (or in more general terms, both the row and the column factors). The GGE biplot is a biplot that displays the GGE of MET data. It is constructed by plotting the first two principal components (PC1 and PC2, also referred to as primary and secondary effects, respectively) derived from singular value decomposition (SVD) of the environment-centered data. Models that decompose the environment-centered data are commonly referred to as sites regression models or SREG, and SREG with two PCs is referred to as $SREG_2$. SREG can be used on scaled or non-scaled data. When replicated data are available, SREG on scaled data (Crossa and Cornelius, 1997) is more desirable because it deals with any heterogeneity of within-site error variance.

One unique merit of a GGE biplot is that it can graphically show the which-won-where patterns of the data, as first described in Yan et al. (2000). Briefly, markers of the cultivars furthest from the plot origin (0,0) are connected with straight lines to form a polygon such that markers of all other cultivars are contained in the polygon. To each side of the polygon, a perpendicular line, starting from the origin of the biplot, is drawn and extended beyond the polygon so that the biplot is divided into several sectors and the markers of the test sites are separated into different sectors. The cultivar at the vertex for each sector is the best performer at sites included in that sector, provided that the GGE is sufficiently approximated by PC1 and PC2. Thus, groups of sites that share the same best performers are graphically identified.

If the which-won-where patterns identified by a biplot are repeatable over years, different mega-environments (subregions) can be defined. By selecting superior cultivars for each mega-environment, both G and GE can be effectively exploited. The GGE biplot is still useful even in cases where the which-won-where patterns are not repeatable over years, which suggests that the tested environments belong to a single mega-environment. It can be used to identify superior cultivars and test environments that facilitate identification of such cultivars, provided that the target mega-environment is sufficiently sampled and that the genotype PC1 scores have near-perfect correlation (say, $r > 0.95$) with the genotype main effects. Ideal cultivars should have large PC1 scores (higher average yield) and near zero PC2 scores (more stable). Similarly, ideal test environments should have large PC1 scores (more discriminating of the cultivars) and near zero PC2 scores (more representative of an average environment). (Note that a "test environment" refers to a year-site combination; it does not necessarily correspond to a "test site".) Thus, the GGE biplot allows many important questions to be addressed effectively and graphically.

However, the requirement for a near-perfect correlation between genotype PC1 scores and genotype main effects is not always met, which restricts to the utility of the $SREG_2$ based GGE biplot. Analysis of the yearly MET data of the Ontario winter wheat performance trials during 1989-1999, and of winter wheat perfor-

W. Yan and L.A. Hunt, Dep. of Plant Agriculture, Univ. of Guelph, Guelph, Ontario, Canada N1G 2W1; P.L. Cornelius, Dep. of Agronomy and Dep. of Statistics, Univ. of Kentucky, Lexington, KY 40546-0091; Jose Crossa, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, 06600 Mexico D.F., Mexico. Received 14 Feb. 2000. *Corresponding author (wyan@uoguelph.ca).

Published in Crop Sci. 41:656–663 (2001).

Abbreviations: G, genotypic main effect; GE, genotype × environment interaction; GGE, Genotype main effects plus genotype × environment interaction; E, environment main effect; $SREG_{M+1}$, Mandel's sites regression model with one additional multiplicative term; PC, principle component; $SREG_2$, Sites regression model with two multiplicative terms; SVD, singular value decomposition.

mance trials from several states of the USA (Yan, unpublished) indicates that the genotype PC1 scores are usually highly correlated with the genotype main effect. Poor correlations between genotype PC1 scores and genotype main effects, however, do occur for some years. Moreover, when multiple years of data are analyzed together, this becomes a norm rather than an exception because of large and complex GE interaction (discussed later). In such cases, the genotype PC1 scores cannot be interpreted as representing the same information as the genotype main effects. Consequently, the yielding ability and stability of the genotypes, and the discriminating ability and the representativeness of the test environments cannot be readily visualized.

To avoid these possible exceptions, in this paper we report an alternative GGE biplot, which is constructed by Mandel's sites regression on genotype main effects as the primary effect and the first principal component derived from subjecting that residual to SVD as the secondary effect. Such a GGE biplot is referred to as a $SREG_{M+1}$ biplot, with the subscript "M" referring to Mandel's solution. In a $SREG_{M+1}$ biplot, the primary effects are the genotype main effects per se; it is, therefore, free from the problem discussed above for the $SREG_2$ biplot. However, it is not clear if a $SREG_{M+1}$ biplot is as effective as the $SREG_2$ biplot in explaining the GGE and in displaying the which-won-where patterns of the data. This study was initiated to answer these questions by comparing the $SREG_2$ biplot and the $SREG_{M+1}$ biplot applied to several datasets that showed different relations between genotype PC1 scores of $SREG_2$ and the genotype main effects.

MATERIALS AND METHODS

The $SREG_2$ Biplot

The $SREG_2$ based GGE biplot is derived from Eq. [1]

$$Y_{ij} - \beta_j = \sum_{n=1}^{2} \lambda_n \xi_{in} \eta_{jn} + \varepsilon_{ij} = \sum_{n=1}^{2} \xi_{in}^{*} \eta_{jn}^{*} + \varepsilon_{ij} \qquad [1]$$

where Y_{ij} is the average yield of Genotype i in Environment j, β_j is the average yield of all genotypes in Environment j, λ_n is the singular value for principal component PCn, ξ_{in} and η_{jn} are scores for Genotype i and Environment j on PCn, respectively, and ε_{ij} is the residual associated with Genotype i in Environment j. The values of λ_n, ξ_{in}, and η_{jn} are simultaneously obtained by subjecting the environment-centered yield (i.e., $Y_{ij}-\beta_j$) to SVD. This can be achieved by principal component analysis of the environment-centered yield using the SAS procedure PRINCOMP. The PRINCOMP generates ξ_{in} as the genotype scores and $(\lambda_n \xi_{in})$ as the environment scores. Alternatively, λ_n, ξ_{in} and η_{jn} can be obtained by the SVD function within the SAS procedure IML, which is a basic function in many SAS procedures related to principal component analysis. A SAS program for principal component analysis of MET data is available from the senior author of this paper.

To display results of fitting Eq. [1] in a biplot, the singular value λ_n has to be absorbed by the singular vector for cultivars h_{jn} and that for environments ξ_{in}. That is, $\xi_{in}^{*} = \lambda_n^{A_n} \xi_{in}$ and $\eta_{in}^{*} = \lambda_n^{1-A_n} \eta_{jn}$. A_n is chosen such that the range of the environment markers is equal to the range of the cultivar markers:

$$\max(\xi_{in}^{*}) - \min(\xi_{in}^{*}) = \max(\eta_{jn}^{*}) - \min(\eta_{jn}^{*}),$$

i.e.,

$$\lambda_n^{A_n}(\max(\xi_{in}) - \min(\xi_{in}) = \lambda_n^{1-A_n}(\max(\eta_{jn}) - (\eta_{jn})).$$

Thus,

$$A_n = 0.5\left\{1 + \frac{\ln\left[\dfrac{\max(\eta_{jn}) - \min(\eta_{jn})}{\max(\xi_{in}) - \min(\xi_{in})}\right]}{\ln\lambda_n}\right\}. \qquad [2]$$

The $SREG_{M+1}$ Biplot

Mandel (1961) presented the following model for analysis of non-additivity of two-way data:

$$Y_{ij} = \beta_j + b_j\alpha_i + \varepsilon_{ij} \qquad [3]$$

where Y_{ij} and β_j are the same as in Eq. [1], α_i is the main effect of Genotype i, and b_j is the regression coefficient of the environment centered yields (i.e., $Y_{ij} - \beta_j$) within Environment j on the genotype main effects (α_i). Equation [3] is similar to the well-known model of Finlay and Wilkinson (1963), but the roles of cultivars and sites are exchanged.

If the first principal component ($\lambda_1\xi_{i1}\eta_{j1}$) from SVD of the residual from Eq. [3], i.e., ($Y_{ij} - \beta_j - b_j\alpha_i$), is added, then

$$Y_{ij} = \beta_j + b_j\alpha_i + \lambda_1\xi_{i1}\eta_{j1} + \varepsilon_{ij} \text{ or}$$

$$Y_{ij} - \beta_j = b_j\alpha_i + \lambda_1\xi_{i1}\eta_{j1} + \varepsilon_{ij} \qquad [4]$$

where all terms are the same as defined in Eq. [1] or [3]. To construct a $SREG_{M+1}$ biplot, Eq. [4] is written as

$$Y_{ij} - \beta_j = b_j^{*}\alpha_i^{*} + \xi_{i1}^{*}\eta_{j1}^{*} + \varepsilon_{ij} \qquad [5]$$

with $\xi_{i1}^{*} = \lambda_1^{A_1}\xi_{i1}$, $\eta_{j1}^{*} = \lambda_1^{1-A_1}\eta_{j1}$, $b_j^{*} = Bb_j$, and $\alpha_i^{*} = B^{-1}\alpha_i$, where A_1 is defined by Eq. [2], and

$$B = \sqrt{\frac{\max(\alpha_i) - (\min(\alpha_i)}{\max(b_j) - \min(b_j)}}. \qquad [6]$$

A_1 and B are chosen such that the plot space used by genotypes are the same as that by environments. Analogous to PC1 and PC2 in the $SREG_2$ model, $b_j^{*}\alpha_i^{*}$ and $\xi_{i1}^{*}\eta_{j1}^{*}$ are referred to as the primary and secondary effects, respectively. All analyses were conducted using SAS (SAS Institute, 1996).

The Data

The data used in this study were from the 1989 to 1999 Ontario winter wheat performance trials (Yan, 1999). Each year, 10 to 33 winter wheat (*Triticum aestivum* L.) cultivars are tested with four to six replicates in seven to 14 sites representing the Ontario winter wheat growing areas. Previous analysis indicated that the yearly variance components due to environment (E) dominated the total yield variation, ranging from 55 to 91% and averaging 80% of the total variance. The variance component due to G ranged from 1.8 to 28.5%, whereas that due to GE ranged from 7.3 to 15.1% (Yan, 1999). G ranged from 13 to 65% of the total GGE. Analysis with the $SREG_2$ biplot revealed that in all years except 1995 the environmental PC1 scores were of the same sign; and in all years except 1995 and 1996 the genotype PC1 scores showed high correlation with the mean yield of the genotypes ($r > 0.93$). Thus, in this study the 1995, 1996, and 1998 datasets, representing different types of relations between genotype PC1 versus genotype main effects, were chosen to compare

Table 1. Proportions of GGE SS explained by $SREG_2$ and $SREG_{M+1}$ for 12 datasets from the 1989–1999 Ontario winter wheat performance trials.

Year	No. of cultivars	No. of sites	Degrees of freedom	% of GGE explained					
				$SREG_2$			$SREG_{M+1}$		
				PC1	PC2	Total	Primary	Secondary	Total
1989	10	9	32	42.5	21.3	63.8	40.7	21.9	62.6
1990	10	7	28	59.7	21.2	80.9	53.5	25.1	78.6
1991	10	9	32	53.3	20.7	74.0	49.1	22.1	71.2
1992	10	10	34	57.0	19.9	76.9	56.4	20.1	76.5
1993	18	9	48	56.8	20.0	76.8	55.4	21.2	76.6
1994	14	11	44	45.6	16.2	61.8	41.6	16.8	58.4
1995	14	14	50	54.2	13.4	67.6	40.8	25.2	66.0
1996	23	9	56	29.6	24.5	54.1	26.7	25.3	52.0
1997	28	8	66	55.0	15.9	70.9	54.0	15.9	69.9
1998	33	8	76	71.5	14.7	86.2	71.0	15.2	86.2
1999	31	9	74	51.5	17.4	68.9	50.7	17.7	68.4
1996–99	11	34	84	24.5	22.7	47.2	23.0	23.9	46.9
Average	–	–	–	50.1	19.0	69.1	46.9	20.9	67.8

the GGE biplot based on $SREG_{M+1}$ with one based on $SREG_2$. In addition, a complete subset of 11 cultivars by 34 environments (year-site combinations) extracted from the 1996 to 1999 trials was also used in the comparison.

RESULTS

For all datasets, both $SREG_2$ and $SREG_{M+1}$ use the same number of degrees of freedom [(g+e−2)+ (g+e−4) or 2(g+e)−6, where g is the number of genotypes and e the number of the environments] (Table 1). With the same number of degrees of freedom, $SREG_2$ is theoretically the most effective model for explaining the variation due to GGE, because the first two principal components are computed to explain the maximum amount of variation. Nevertheless, $SREG_{M+1}$ explained only slightly smaller amounts of GGE. When averaged over 12 datasets, $SREG_2$ explained 69.1%, whereas $SREG_{M+1}$ explained 67.8% of the total GGE (Table 1). Thus, $SREG_{M+1}$ is nearly as effective as $SREG_2$ in explaining the variation of GGE. So the discussion will be focused on whether the $SREG_{M+1}$ biplot displays similar which-won-where pattern as the $SREG_2$ biplot.

1998 Data

The PC1 scores of the $SREG_2$ model had near-perfect correlation (r = 0.99) with the genotypic main effects for this dataset. Consequently, the $SREG_2$ biplot and the $SREG_{M+1}$ biplot look almost exactly alike. They were, therefore, equally effective in displaying the GGE information (Fig. 1A and 1B).

The GGE biplot is constructed by plotting the primary effect scores of each genotype (as x-axis) and each site against their respective secondary effect scores (as y-axis) such that each genotype and each test site is represented by a "marker." For visualizing the which-won-where pattern, the genotype markers located away from the plot origin were first visually identified and connected with straight lines to form a polygon, within which the markers of all other genotypes are contained. These away-from-origin genotypes, namely 6, 9, 29, 33, 27, 28, 20, and 2 in Fig. 1A, are called "corner" or "vertex" genotypes because they are at the corners of the polygon. Next, starting from the origin, lines perpendicular to the sides of the polygon are drawn to, and extended beyond, each side of the polygon dividing the plot into several sectors; each site will fall into one of the sectors (note that only perpendiculars relevant to discussion were drawn). Assuming that the biplot sufficiently approximates the variation of GGE, it can be mathematically proven that all sites in the same sector share the same winning genotype, which is the vertex genotype for that sector (Yan et al., 2000).

In Fig. 1A, the sites fell into three sectors: the winning genotype for sites RN, WE, ID, and NN was Genotype 6; the winning genotype for sites WK, HN, and EA was Genotype 9; and the winning genotype for site OA was Genotype 29. Note that Genotype 9 was the best performer for WK, HN, and EA because markers of these sites were on Genotype 9's side of the perpendicular to the line that connects Genotypes 9's marker and that of genotype 6. Vertex genotypes without any site in their sectors were not the highest yielding genotypes at any site; moreover, they were the poorest genotypes at all or some sites. Genotypes within the polygon, particularly those located near the plot origin, were less responsive than the vertex genotypes. It can be appreciated that the supplementary lines on the biplot are critical for visual analysis of the MET data.

In addition, a near-perfect correlation between genotype primary effect scores and the genotype main effects allows both biplots, Fig. 1A, as well as Fig. 1B, to be used to evaluate cultivars for their yielding ability and stability and to evaluate environments for their discriminating ability and representiveness. Genotypes 6 and 9 gave the highest average yields (largest primary scores) and were relatively stable over the sites (small absolute secondary scores). In contrast, three non-adapted Genotypes 27, 28, and 31 yielded poorly at all sites, as indicated by their small primary scores (low yielding) and relatively small secondary scores (relatively stable). The average yield of Cultivars 1 and 20 were below average (primary scores <0) and highly unstable (large absolute secondary values). The biplots show not only the average yield of a genotype (the primary effect), but also how it was achieved. That is, the biplots also show the yield of a genotype at individual sites. For example,

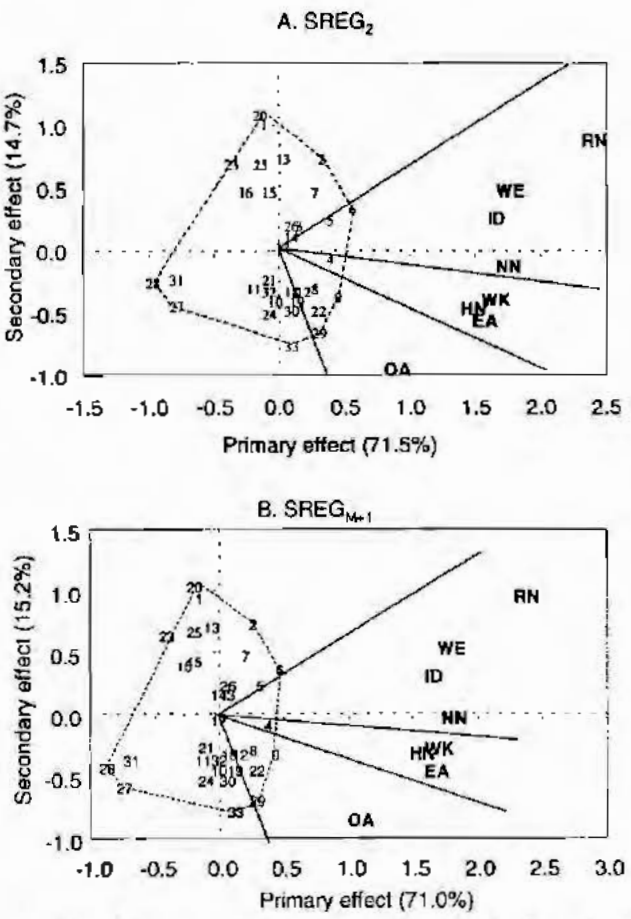

Fig. 1. $SREG_2$ and $SREG_{M+1}$ biplots for the 1998 Ontario winter wheat performance trial data. The numbers are different cultivars; the sites are EA = Elora, HN = Harriston, ID = Inwood, NN = Nairn, OA = Ottawa, RN = Ridgetown, WE = Woodslee, WK = Woodstock.

Cultivar 6 had the highest average yield because it yielded the highest at sites RN, WE, ID, and NN, and yielded above average at all other sites. On the other hand, the average yield of Cultivar 20 was below average, because it yielded below average at sites OA, EA, HN, WK, and NN, even though it was quite good at RN. A below-average yield is indicated if the virtual line from the origin to the marker of a genotype has an obtuse angle with the virtual line from the origin to the marker of a test site. Likewise, an above-average yield is indicated by an acute angle. Supplementary lines, not presented in the biplots, are required to explicitly determine these relationships.

With respect to the test sites, RN was most discriminating as indicated by the longest distance between its marker and the origin. However, due to its large secondary score, cultivar differences observed at RN may not exactly reflect the cultivar differences in average yield over all sites. Site NN was not the most discriminating, but cultivar differences at NN should be highly consistent with those averaged over sites because it had a near-zero secondary effect score. At a site with a near-zero secondary effect score, the genotypes are essentially ranked according to their primary effect scores (i.e., genotype main effects since they were perfectly correlated in this dataset) and the differences among genotypes are in proportion to the primary effect scores of the sites. Thus, a genotype that yielded well at such a site has a large average yield. On the contrary, site OA was neither discriminating (small primary effect score) nor representative (large secondary effect score); and therefore, cultivars had high yield at OA did not necessarily give high average yield over sites. Analysis of multiple year data indicated that OA represented a different mega-environment (eastern Ontario) from the major winter wheat growing regions in Ontario (Yan et al., 2000; Yan, 1999).

1996 Data

As with most datasets, the $SREG_2$ biplot (Fig. 2A) for 1996 indicates that all PC1 scores of the sites were of the same sign, which was arbitrarily assigned positive so that the genotype PC1 scores correlated positively with the genotype main effect. However, as mentioned earlier, the correlation between the genotype PC1 scores and the genotype main effects for this dataset was only 0.85. The relatively poor correlation is associated with the fact that the GGE explained by PC1 is only slightly greater than that by PC2 (29.6 vs. 24.5%). The poor correlation prevents the genotype PC1 scores of the $SREG_2$ solution being interpreted as representing the genotype main effect; in fact, it alone is not interpretable in known biological and agricultural terms. In such cases, the utility of a $SREG_2$ biplot is limited to investigation of the which-won-where patterns. Based on Fig. 2A, Cultivar 1 was the best performer at sites RN, LN, ID, and WE; and Cultivar 2 was the best performer at sites EA, WK, CA, and OA, and nearly the best at HW.

The $SREG_{M+1}$ biplot (Fig. 2B) explained slightly less GGE, but revealed the same which-won-where patterns as the $SREG_2$ biplot. It indicates that Cultivar 1 won at sites RN, LN, WE, and ID, and Cultivar 2 won at sites EA, WK, CA, HW, and OA. In addition, the $SREG_{M+1}$ biplot is more interpretable. By definition, the primary effects of the $SREG_{M+1}$ biplot are the cultivar main effects, and its secondary effects are deviations from the main effects of the cultivars. Thus, the $SREG_{M+1}$ biplot explicitly showed that Cultivars 1 and 2 were the highest yielding cultivars on average, but neither was very stable, as evidenced by their relatively large secondary effects. With respect to the sites, the $SREG_{M+1}$ biplot indicated that site EA was highly discriminating, but not representative of the average environment, whereas WK and RN were both discriminating and representative.

1995 Data

The 1995 dataset was the only dataset found during the 1989 to 1999 Ontario winter wheat performance trials in which the site PC1 scores of the $SREG_2$ differ in sign (Fig. 3A). Among the 14 test sites, four (Sites 4, 6, 7, and 10) had negative PC1 scores, though their absolute values were small. This led to poor a correlation between the cultivar PC1 scores and the cultivar

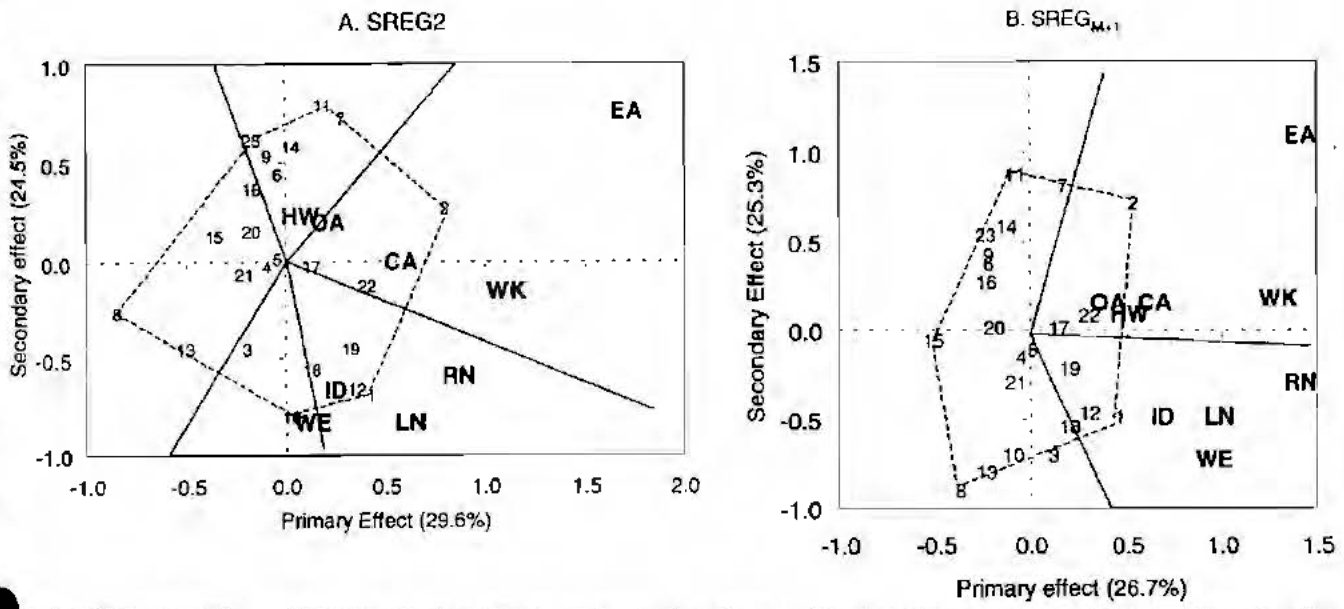

Fig. 2. SREG$_2$ and SREG$_{M+1}$ biplots for the 1996 Ontario winter wheat performance trial data. The numbers are different cultivars; the sites are CA = Centralia, EA = Elora, HN = Harriston, HW = Harrow, ID = Inwood, OA = Ottawa, RN = Ridgetown, WE = Woodslee, WK = Woodstock.

main effects ($r = 0.83$). The SREG$_2$ biplot indicates that cultivar G6 was the best for nearly all sites except Sites 4, 6, and 7, at which Cultivar G4 (and also G10) was better than G6. Cultivar G7 was as good as G6 for Sites 5 and 12. These patterns are similar in the SREG$_{M+1}$ biplot (Fig. 3B). It indicates that Cultivar G6 was on average the best and Cultivar G12 the second best, and that Sites 5 and 12 were highly discriminating but neither was representative. Interestingly, all sites had positive primary effects in the SREG$_{M+1}$ biplot, as compared with the site PC1 scores of different signs in the SREG$_2$ biplot.

1996–1999 Data

Although the environmental PC1 scores in the SREG$_2$ model tend to be of the same sign for yearly MET, they often take different signs when multi-year data are jointly analyzed. For this dataset, among all 34 year-site combinations, 9 had negative PC1 scores and the rest had positive PC1 scores (Fig. 4A). Like the 1996 data, the GGE explained by PC1 was only slightly greater than that by PC2 (24.5 vs. 22.7%). As a result, the correlation between cultivar PC1 scores and cultivar main effects was only 0.58. This low correlation prevents

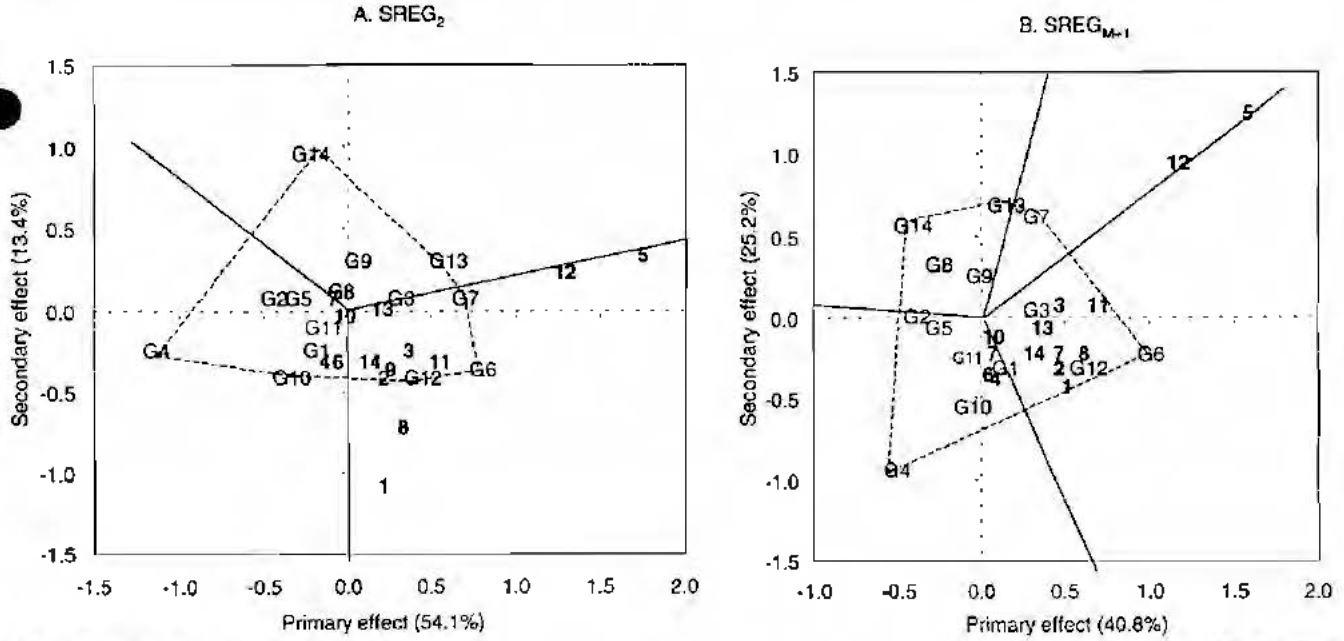

Fig. 3. SREG$_2$ and SREG$_{M+1}$ biplots for the 1996 Ontario winter wheat performance trial data. Each site is represented by a number, and each cultivar is represented by a number preceded by the letter G.

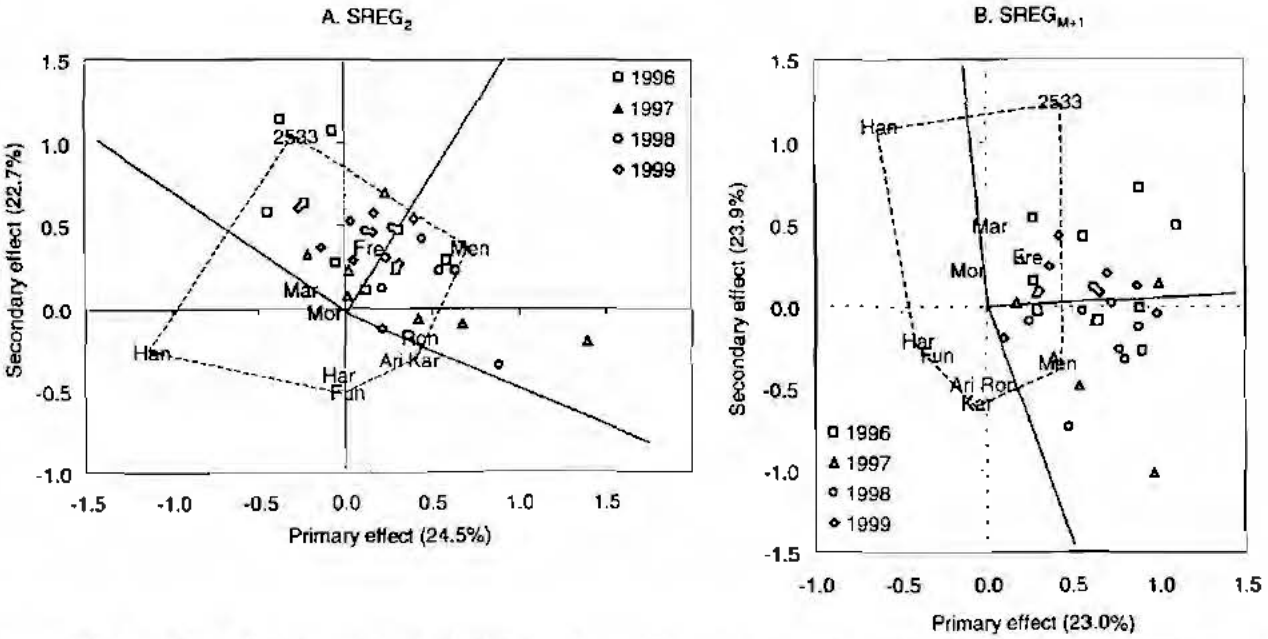

Fig. 4. $SREG_2$ and $SREG_{M+1}$ biplots for the 1996–1999 Ontario winter wheat performance trial data. Sites in different years are represented by different symbols. The full cultivar names are: 2533 = Pioneer 25W33, Ari = OAC Ariss, Fre = Freedom, Fun = Fundulea, Han = Hanover, Har = Harus, Kar = Karena, Mar= Marilee, Men = Mendon, Mor = AC Morley, Ron = AC Ron.

visual identification of cultivars with high average yield based on the $SREG_2$ biplot. Nevertheless, as with all previous datasets, both biplots displayed very similar which-won-where patterns (Fig. 4A and 4B). The $SREG_2$ biplot predicted that cultivar "2533" was the best performer in about half of the 34 environments while cultivar "Men" was the best in the other half. Therefore, it can be inferred that cultivars "2533" and "Men" must be the two best performers on average. This, however, is explicitly indicated only in the $SREG_{M+1}$ biplot. As for the 1995 dataset, while the primary effects of the environments were of different signs in the $SREG_2$ biplot, they were all positive in the $SREG_{M+1}$ biplot.

DISCUSSION

Merits of the Two Types of GGE Biplots

This study indicates that both the $SREG_2$ biplot and the $SREG_{M+1}$ biplot explained similar amounts of variation due to GGE, although the former tends to explain slightly more in most cases. Both biplots displayed the same which-won-where pattern and indicated the same winning cultivars in individual environments. Therefore, the two biplots can be considered as equally effective in these regards.

The $SREG_{M+1}$ biplot was designed to be more interpretable than the $SREG_2$ biplot. First, since the genotypic scores for the primary effect of $SREG_{M+1}$ are designated to indicate the average yield (general adaptation) of the cultivars, the genotypic scores of the secondary effect must indicate GE interaction associated the cultivars, which is an indicator of selective or specific adaptation. Thus, the $SREG_{M+1}$ biplot simultaneously displays both general adaptation and specific adaptation (stability) of the cultivars. The ideal cultivars are those with large primary effect scores but near-zero secondary scores. Second, because the genotypic primary effects indicate general adaptation of the cultivars, the environmental primary effects must indicate the ability of the environments to discriminate among the cultivars in terms of general adaptation. Environments with larger primary effects would thus facilitate identification of cultivars with better general adaptation. Third, analogous to the genotypic secondary effects, the environmental secondary effects must indicate the tendency of each environment to cause GE interaction. Environments with large (absolute) secondary effects should favor the performance of some cultivars, but disfavor others at the same time. Thus, cultivars selected under environments with large secondary effects may be highly specific to these environments but lack general adaptation or stability. Therefore, from the perspective of selection for high yielding and stable cultivars, the ideal test environments should have large primary effects, but near-zero secondary effects.

Why Correlation between Genotype Scores of PC1 in $SREG_2$ and Genotype Main Effects Varies with Datasets

It was concluded that the $SREG_{M+1}$ biplot is more desirable than the $SREG_2$ biplot for MET data analysis because the interpretability of the latter is impacted by the uncertain relations between its primary effects and the genotype main effects. On the basis of the trials investigated in this study, Fig. 5 indicates that this correlation is strongly determined by the relative importance of G in GGE. Near-perfect correlation occurs when G is 40% or more of GGE (the 1992, 1993, 1997–1999 datasets), and poor correlation occurs when G is 20% or less of GGE (the 1995, 1996 and 1996–1999 datasets).

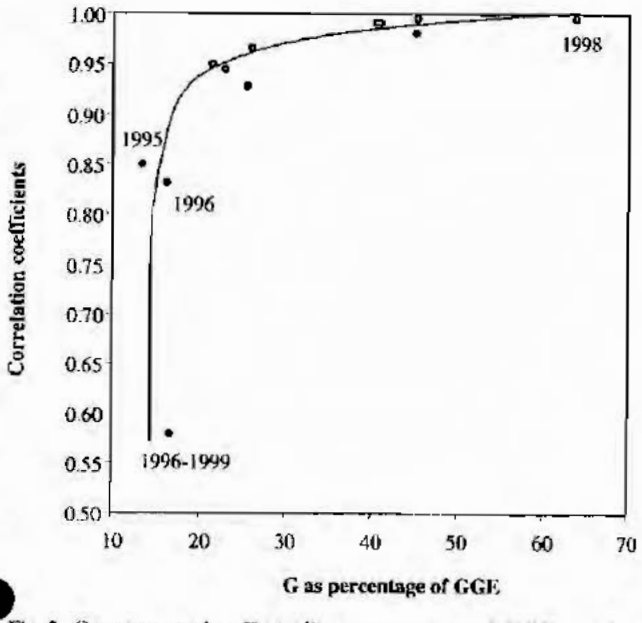

Fig. 5. Genotype main effect (G) as percentage of GGE and the correlation coefficient (r) between the genotype PC1 scores of the $SREG_2$ model and the genotype main effects for 12 datasets from the 1989–1999 Ontario winter wheat performance trials.

The essence of principal component analysis is to pick up the most important pattern in the data using the smallest number of degrees of freedom. PC1 picks up the largest pattern, PC2 picks up the second largest pattern, and so on. A close correlation between PC1 scores and genotype main effects occurs only when the genotype main effect is large enough to be the most important component of GGE. A poor correlation occurs otherwise, which suggests strong and complex GE interaction in the data. Therefore, it is not surprising that the correlation between PC1 scores of $SREG_2$ and genotype main effect is typically poor when multi-year data are analyzed in a genotype × environment (year-site) fashion, because greater and more complex GE interactions are sampled in a multi-year MET than in a single year MET. Complex GE interaction is usually accompanied by similar amounts of GGE explained by PC1 and PC2 (as for the 1996 and 1996–1999 datasets, Table 1), as opposed to much more GGE explained by PC1 than by PC2 (e.g., the 1998 dataset).

The Usefulness of the GGE Biplot Based on a Single Year MET

As a graphic approach to MET data analysis, GGE biplot can be useful in two major aspects. The first is to display the which-won-where pattern of the data, which may lead to identification of different mega-environments. The second is to identify high-yielding and stable cultivars and discriminating and representative test environments. However, both promises are based on the assumption that the data is sufficiently representative of the target environment; a conclusion can never go beyond what the data allow. While multi-year MET data are required for any decisive cultivar and site evaluation, they are normally unbalanced, and therefore the biplot technique can not readily applied; single year data are usually balanced but they may not be representative of future years. Thus, a question arises whether biplot analysis of single year MET data is really useful if the which-won-where pattern is not repeatable over years.

A single year data may indeed have limited value because of the year-to-year variation. Nevertheless, we believe biplot analysis of single year MET data is worthwhile for the following reasons. First, the GGE biplot is a graphic display of the G and GE of the data, which are relevant to cultivar evaluation and mega-environment identification. Therefore, if the researcher believes that a single year MET is worthy of analysis, and we believe most researchers do, the GGE biplot technique should be the first choice. Although the biplot does not add new information to the data, it does help the researcher quickly view the patterns that are in the data. The biplot gives the researcher the power to "see" what was going on in a particular year. Some may question the usefulness of the single year patterns if they are not repeatable over years. But without knowing the patterns from individual years, how could one know if they are repeatable or not? Second, the biplot can be used to identify research problems. For example, if two cultivars were found to perform the best in two different groups of locations in a particular year, one might want to know what were the underlying reasons, and answers to this question may lead to valuable findings. By relating biplot scores to explanatory variables collected in the trials, Yan and Hunt (2001) was able to reveal that in Ontario, Canada, tall and late winter wheat cultivars tended to be favored in seasons with cold winters and cool summers, whereas early and short cultivars tended to be favored in seasons with warm winters and hot summers. Third, the biplot patterns based on a single year MET can serve as hypotheses, which can be tested using extended data and more critical statistics. For example, biplots based on yearly data from the Ontario winter wheat performance trials led to the hypothesis that two eastern Ontario sites (Ottawa and Kemptville) constituted a mega-environment different from the rest of the Ontario winter wheat growing region, which was subsequently tested and supported by variance component analysis based on pooled data from 11 yr of performance trials (Yan, 1999). Thus, although conclusions from a single year MET may not be decisive, they are valuable suggestions. Fourth, even if the which-won-where pattern is proven to be unrepeatable over years, the researcher would still want to know the average yield and the stability of the cultivars based on each year's MET. These two aspects of cultivar performance are graphically depicted by the abscissa and ordinate of the biplot, respectively. Finally, although a biplot from a single year may not be very informative, biplots constructed from several years can be highly valuable.

Moreover, the biplot technique is not limited to single year MET data analysis. It can also be applied to balanced subsets extracted from multiple years of trials. In Ontario, for example, over 20 winter wheat cultivars are common to three to four years of performance trials,

and a balanced subset from such database should contain valuable information. Furthermore, the biplot technique is not even limited to genotype × environment data analysis. It can also be used in displaying and analyzing other types of two-way data such as genotype × trait data and diallel cross data (Yan, unpublished research). In conclusion, the GGE biplot is a useful tool for, but not limited to, MET data analysis.

REFERENCES

Crossa, J., and P.L. Cornelius. 1997. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. Crop Sci. 37:405–415.

Finlay, K.W., and Wilkinson, G.N. 1963. The analysis of adaptation in a plant breeding program. Aust. J. Agric. Res. 14:742–754.

Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453–467.

Mandel, J. 1961. Non-additivity in two-way analysis of variance. J. Am. Stat. Assoc. 65:878–888.

SAS institute, 1996. SAS/STAT user's guide, second edition. SAS institute Inc., Cary, NC.

Yan, W. 1999. A study on the methodology of yield trial data analysis—with special reference to winter wheat in Ontario. Ph D diss., University of Guelph, Guelph, Ontario, Canada.

Yan, W., L.A. Hunt, Q., Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci. 40:597–605.

Yan, W., and L.A. Hunt. 2001. Genetic and environmental causes of genotype × environment interaction for winter wheat yield in Ontario. Crop Sci. 41:19–25.

Interpretation of Genotype × Environment Interactions for Early Maize Hybrids over 12 Years

C. Epinat-Le Signor, S. Dousse, J. Lorgeou, J.-B. Denis, R. Bonhomme, P. Carolo, and A. Charcosset*

ABSTRACT

Genotype × environment interaction was investigated for grain yield of early maize (Zea mays L.) hybrids. Data were obtained from the French Association Générale des Producteurs de Maïs trial network and included 132 hybrids and 229 environments over 12 yr, following an unbalanced design. Analysis of genotype × environment interaction was done for the 1-yr data sets, for the two successive years data sets, and for the 12-yr data set. The magnitude of genotype × environment interaction variance was equal to, or greater than the genotypic variance. Interaction effect was modeled by factorial regression analysis using additional genotypic and environmental information. Genotypic covariates considered were the sum of growing day degrees (GDD) necessary from sowing to flowering and the GDD necessary from flowering to maturity. Environmental covariates were the mean temperature from sowing to the 12 leaf stage, the mean temperature from the 12 leaf stage to the end of the linear grain-filling stage, the water balance around flowering, and the sum of solar radiation around flowering. These six covariates explained about 40% of the interaction effect in all analyses, with equal contribution of genotypic variates (20%) and environmental variates (20%). Flowering earliness of hybrids, water balance around flowering, and mean temperature from the 12 leaf stage to the end of the grain filling phase were determinants of genotype × environment interaction for grain yield in the considered area. A biological interpretation of the interaction was attempted through examination of the regression parameters.

NEWLY REGISTERED CULTIVARS generally need to be tested at many locations and for several years before being recommended for a given zone. To achieve this goal, multi-environment trials form the core of varietal testing programs in many countries. These programs have to face the recurring problem of genotype × environment (GE) interactions. Indeed, differential genotypic responses to variable environmental conditions, especially when associated with changes in genotypic ranking, limit the identification of superior, stable hybrids. The GE interactions are as much a function of the environmental variables as a function of the genotypic, morphological, phenological, and physiological traits of the varieties (Nachit et al., 1992). Identification of causal factors of the GE effect and quantification of unexplained variation are of prime importance for selecting for stability or to recommend environmentally specific varieties. During recent decades, new developments have been achieved in crop physiology, agronomy, and statistics and some integrated approaches appeared for GE interactions evaluation (Brancourt-Hulmel, 1999). Many fixed or mixed models have been used for detecting and characterizing GE interaction (van Eeuwick, 1995a,b; Yan and Hunt, 1998; Vargas et al., 1999).

Until now, there have been few attempts to analyze this interaction for the newly registered varieties of maize over an important series of years. Only van Eeuwick et al. (1995b) reported results concerning maize multi-environment trials over a series of 11 yr but they studied forage percent dry-matter content and not yield. Little is known about the most relevant environmental

C. Epinat-Le Signor, S. Dousse, A. Charcosset, INRA-INAPG-UPS, Station de Génétique Végétale, Ferme du Moulon, F91190 Gif sur Yvette; J. Lorgeou, Association Générale des Producteurs de Maïs, Station expérimentale, F91720 Boigneville; J.-B. Denis, INRA, Station de Biométrie, Route de Saint-Cyr, F78026 Versailles; R. Bonhomme, INRA, Unité de recherche en Bioclimatologie, F78850 Thiverval-Grignon; P. Carolo, Rustica-Prograin Génétique, 117 avenue de Vendôme, F41000 BLOIS. Received 18 Aug. 1999. *Corresponding author (charcos@moulon.inra.fr).

Published in Crop Sci. 41:663–669 (2001).

Abbreviations: AGPM, Association Générale des Producteurs de Maïs, France; AMMI, Additive Main effects and Multiplicative Interaction analysis; GDD, sum of Growing Day Degrees; GDDs_f, GDD from sowing to flowering; GDDf_m, GDD from flowering to maturity; GE, genotype × environment; Mg ha^{-1}, ton per hectare; RSD, Residual Standard Deviation; SRf, sum of radiation around flowering (from 06–20 to 08–20); SS, Sum of Squares; TMs_12l, mean temperature from sowing to 500 GDD (12 leaves stage); TM12l_e, mean temperature from 500 GDD to 1425 GDD (end of linear grain-filling phase); WATf, water balance around flowering (rainfall + irrigation – evapotranspiration from 06–20 to 08–20).

CROP BREEDING, GENETICS & CYTOLOGY

Using Partial Least Squares Regression, Factorial Regression, and AMMI Models for Interpreting Genotype × Environment Interaction

Mateo Vargas, José Crossa,* Fred A. van Eeuwijk, Martha E. Ramírez, and Ken Sayre

ABSTRACT

Partial least squares (PLS) and factorial regression (FR) are statistical models that incorporate external environmental and/or cultivar variables for studying and interpreting genotype × environment interaction (GEI). The Additive Main effect and Multiplicative Interaction (AMMI) model uses only the phenotypic response variable of interest; however, if information on external environmental (or genotypic) variables is available, this can be regressed on the environmental (or genotypic) scores estimated from AMMI and superimposed on the AMMI biplot. The objectives of this study with two wheat [*Triticum turgidum* (L.) var. *durum*] field trials were (i) to compare the results of PLS, FR, and AMMI on the basis of external environmental (and cultivar) variables, (ii) to examine whether procedures based on PLS, FR, and AMMI identify the same or a different subset of cultivar and/or environmental covariables that influence GEI for grain yield, and (iii) to find multiple FR models that include environmental and cultivar covariables and their cross products that explain a large proportion of GEI with relatively few degrees of freedom. Results for the first trial showed that AMMI, PLS, and FR identified similar cultivar and environmental variables that explained a large proportion of the cultivar × year interaction. Results for the second wheat trial showed good correspondence between PLS and FR for 23 environmental covariables. For both trials, PLS and FR complement each other and the AMMI and PLS biplots offered similar interpretations of the GEI. The FR analysis can be used to confirm these results and to obtain even more parsimonious descriptions of the GEI.

MULTI-ENVIRONMENT TRIALS play an important role in selecting the best cultivars (or agronomic practices) to be used in future years at different locations and in assessing a cultivar's stability across environments before its commercial release. When the performance of cultivars is compared across sites, several cultivar attributes are considered, of which grain yield is one of the most important. Cultivars grown in multi-environment trials react differently to environmental changes. This differential response of cultivars from one environment to another is called genotype × environment interaction (GEI).

Genotype × environment interaction has been studied, described, and interpreted by means of several statistical models (Crossa, 1990). Some models, such as analysis of variance, regression on the environmental mean models (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966), and the Additive Main effects and Multiplicative Interaction (AMMI) models (Gollob, 1968; Mandel, 1971; Kempton, 1984; Gauch, 1988) use only the phenotypic response variable of interest. The AMMI model is more parsimonious than the conventional analysis of variance model in describing GEI and provides greater scope for modeling and interpreting GEI than the simple regression on the site mean because GEI can be modeled in more than one dimension.

When information on external environmental (or genotypic) variables such as meteorological data or soil information is available, these variables can be correlated to or regressed on the environmental (genotypic) scores estimated by AMMI. Information from these regressions can be superimposed on the AMMI biplot along with cultivar and environmental scores (van Eeuwijk, 1995), so that interpretation of the grain yield GEI is possible. External environmental information cannot be used directly in the AMMI model, however. When additional information on external cultivar variables is available (physiology, maturity, disease reaction, genetic markers, etc.), other statistical models, including factorial regression models (FR) (Denis, 1988; van Eeuwijk et al., 1996) and partial least squares (PLS) regression (Aastveit and Martens, 1986; Talbot and Wheelwright, 1989; Vargas et al., 1998) can be used to determine which of these external environmental or cultivar variables influence GEI of grain yield.

Factorial regression models are ordinary linear models that explain GEI by differential cultivar sensitivity to explicit external environmental variables (environmental characterization) and have the advantage that hypotheses about the influence of those external variables on GEI of grain yield can be statistically tested. As with all linear regression models, factorial regression models become difficult to deal with when there are many explanatory variables that are highly correlated—the multi-collinearity problem. PLS regression models are appropriate for these situations. As in factorial regression, PLS regression describes GEI in terms of dif-

M. Vargas, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática (ISEI), Colegio de Postgraduados, CP 56230, Montecillo, Mexico, Universidad Autónoma Chapingo, CP 56230, Chapingo, Mexico, and International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; J. Crossa, Biometrics and Statistics Unit, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; F.A. van Eeuwijk, Dep. of Agricultural, Environmental and Systems Technology, Wageningen Agricultural Univ., Dreijenlaan 4, 6703 HA Wageningen, the Netherlands; M.E. Ramírez, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática (ISEI), Colegio de Postgraduados, CP 56230, Montecillo, Mexico; Ken Sayre, Wheat Program, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico. Received 2 July 1998. *Corresponding author (JCROSSA@CIMMYT.MX).

Published in Crop Sci. 39:955–967 (1999).

Abbreviations: AMMI, Additive Main effect and Multiplicative Interaction; GEI, genotype × environment interaction; FR, factorial regression; PLS, partial least squares.

ferential sensitivity of cultivars to environmental variables. The difference is that the explanatory variables are hypothetical, synthetic variables (linear combinations of the complete set of measured environmental and/or cultivar variables) and there is no limit to the number of explanatory covariables that can be used. The PLS regression models are not linear models, so standard linear regression theory for testing cannot be used; however, good alternatives are available.

The advantages and/or disadvantages of the above mentioned statistical models for studying and interpreting GEI with a large number of external environmental and/or cultivar variables have not been compared. Therefore, the objectives of this study were to: (i) compare the results from AMMI, PLS, and FR in two wheat trials when a large set of external environmental (and cultivar) covariables are available, (ii) examine whether procedures based on AMMI, FR, and PLS identify the same or different subsets of cultivar and/or environmental covariables that explain GEI for grain yield, and (iii) find more parsimonious multiple FR models that include environmental and cultivar covariables and their cross products that explain large proportion of GEI with relatively few degrees of freedom.

MATERIALS AND METHODS

Theory

van Eeuwijk (1996) gave a comprehensive description of the AMMI and FR models and how to apply them to assess, study, and interpret GEI. Vargas et al. (1998) described the theory of PLS in the context of GEI and detailed its algorithm. The AMMI, FR, and PLS models are briefly described here.

AMMI Model and the Biplot. A basic model for the analysis of the two-way table of cultivar yield by environment data is the analysis of variance model:

$$E(y_{ij}) = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} \qquad [1]$$

where E stands for expectation, μ is the grand mean, τ_i is the main effect of the ith cultivar, β_j is the main effect of the jth site, and $(\tau\beta)_{ij}$ is the GEI effect of the ith cultivar in the jth environment.

Model [1] can be written in matrix notation as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \boldsymbol{\tau\beta} \qquad [2]$$

where $\mathbf{Y} = (y_{ij})$ is the data matrix of size $I\times J$ of grain yield of I cultivars in J environments, μ is a scalar representing the grand mean, $\boldsymbol{\tau} = (\tau_i)$ is a $I\times 1$ vector of main effects of cultivars, $\boldsymbol{\beta} = (\beta_j)$ is a $J\times 1$ vector of main effects of sites, and $\boldsymbol{\tau\beta} = (\tau\beta)_{ij}$ is the $I\times J$ interaction matrix (not a vector product) where each element of the matrix specifies the interaction effect for the ith cultivar in the jth site. $\mathbf{1}_I$ and $\mathbf{1}_J$ are unit vectors of size $I\times 1$ and $J\times 1$, respectively. The common constraints are $\mathbf{1}_I'\boldsymbol{\tau} = \mathbf{1}_I'\boldsymbol{\beta} = 0$ and $\mathbf{1}_I'\boldsymbol{\tau\beta}\,\mathbf{1}_J' = 0$.

As mentioned previously, a commonly used procedure for modeling GEI is the simple regression of cultivar performance on the environment mean such that $(\tau\beta)_{ij} = \zeta_i\beta_j + d_{ij}$, where ζ_i measures the sensitivity of the ith cultivar to prevailing environmental conditions in the multiplicative (bilinear) term $\zeta_i\beta_j$ and d_{ij} is the residual term (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966). This model can be depicted as a set of straight lines with different slopes, one for each cultivar, where the heterogeneity of slopes accounts for the GEI. Since heterogeneity of slopes in this model generally explains only a small proportion of the usually complex GEI, a more elaborate model is often necessary for an adequate description of GEI.

A generalization of the regression on the site mean model is the multiplicative (bilinear) model

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \textstyle\sum_{k=1,K}\lambda_k\,\theta_{ik}\,\gamma_{jk} \qquad [3]$$

also called Principal Component Analysis (PCA) of the GEI or Additive Main effect and Multiplicative Interaction (AMMI) model (Gollob, 1968; Mandel, 1971; Kempton, 1984; Gauch, 1988) or biadditive model (Denis and Gower, 1994). The parameters μ, τ_i, and β_j are the same as in the analysis of variance model. K being the number of multiplicative (bilinear) terms in the model. The λ_k are scaling constants obtained from the singular value decomposition of the residual matrix consisting of the two-way table of means corrected for cultivar and site main effects (residual from additivity), $w_{ij} = \bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}$ (Gabriel, 1978) and are ordered such that $\lambda_k \geq \lambda_{k+1}$. The θ_{ik} are cultivar interaction parameters (or scores) that measure cultivar sensitivity to hypothetical environmental factors denoted by environmental interaction parameters γ_{jk} (or scores). Orthonormality constraints for the cultivar and environmental scores are $\Sigma_i\,\theta_{ik}^2 = \Sigma_j\,\gamma_{jk}^2 = 1$ and $\Sigma_i\theta_{ik}\,\theta_{ik'} = \Sigma_j\,\gamma_{jk}\,\gamma_{jk'} = 0$ for $k \neq k'$.

For determining the number of multiplicative terms to be retained in a multiplicative model, various tests can be used. The F-test of Gollob (1968) uses the ratio between the mean square for axis k against an estimate of the error term. The mean squares of axis k is calculated by taking the square of λ_k, and the corresponding degrees of freedom, computed by $(I - 1) + (J - 1) - (2k - 1)$. Using simulation studies, Cornelius (1993) showed that the Gollob's F-test is very liberal. Thus, in this study we have used the approximate F-tests F_{GH1}, F_R, and F_1 (Cornelius et al., 1993, 1996; Cornelius, 1993).

Model [3] written in matrix notation is

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \boldsymbol{\Theta}\,\boldsymbol{\Lambda}\,\boldsymbol{\Gamma}' \qquad [4]$$

where the first three terms on the right side are the same as in Eq. [2]. The fourth term represents the GEI, where $\boldsymbol{\Theta} = (\theta_{ik})$ is a $I\times K$ matrix , $\boldsymbol{\Lambda} = (\lambda_{kk})$ is a $K\times K$ diagonal matrix and $\boldsymbol{\Gamma} = (\gamma_{jk})$ is a $J\times K$ matrix. The normalization and orthogonality constraints are $\mathbf{1}_I'\boldsymbol{\tau} = \mathbf{1}_J'\,\boldsymbol{\beta} = 0$, $\mathbf{1}_I'\boldsymbol{\Theta} = \mathbf{1}_J'\,\boldsymbol{\Gamma} = \mathbf{0}$, where $\mathbf{0}$ is a matrix of zeros of size $I\times K$, and $\boldsymbol{\Theta}'\boldsymbol{\Theta} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{I}_K$. The kth bilinear term, $k = 1, \ldots, K$, is formed by a score θ_{ik} specific to Cultivar i, a scale constant factor λ_{kk} and a score γ_{jk} specific to Site j.

The results of AMMI analysis can be presented graphically in the form of biplots (Gabriel, 1971) in which the cultivar and environment scores of the first two or three bilinear (multiplicative) terms are represented by vectors in a space, with starting points at the origin and end points determined by the scores. Usually the environmental and cultivar scores of the first and second bilinear terms are plotted. The distance between two cultivar vectors (their end points) is indicative of the amount of interaction between the cultivars. The cosine of the angle between two cultivar (or environment) vectors approximates the correlation between the cultivars (or environments) with respect to their interaction. Acute angles indicate positive correlation, with parallel vectors (in exactly the same directions) representing a correlation of 1. Obtuse angles represent negative correlations, with opposite directions indicating a correlation of −1. Perpendicularity of directions indicates a correlation of 0. The relative amounts of interaction for a particular cultivar over environments can be obtained from orthogonal projections of the environmental vectors on the line determined by the direction of the corresponding cultivar vector. Environmental vectors having the same direction as the cultivar vectors have positive interactions (that is, these environments favored these cultivars), whereas vectors in the opposite direction have negative interactions.

The biplot obtained from AMMI can be enriched by a

procedure described by van Eeuwijk (1995). Information on external environmental (or cultivar) variables can be correlated to or regressed on the environmental or cultivar interaction parameters (θ_{ik} or γ_{jk}) estimated from AMMI and incorporated into the biplot so that a better interpretation of the GEI of grain yield can be attempted. Once it has been decided that the AMMI solution has, for example, two axes for interaction, the squared correlation coefficients from the regressions of the covariables on the scores of both axes simultaneously (regression through the origin) are computed. When this squared correlation is sufficiently high, information for the covariables can be drawn in the biplot by giving them a direction that is determined by the regression coefficients (van Eeuwijk, 1995). For example, if one environmental covariable is regressed on the AMMI environmental scores of Axes 1 and 2, then coefficients b_{axis1} and b_{axis2} serve as coordinates for that covariable in the biplot. When the environments are projected in the direction determined by the regression coefficients (b_{axis1}, b_{axis2}), it gives the rank-order of the environments on this environmental covariable. The same can also be done for cultivar covariables.

Factorial Regression Models. Factorial Regression models also have multiplicative structure for the interaction, like the AMMI model. The main difference between FR models and multiplicative models such as AMMI model is that in FR the GEI (residual matrix consisting of the two-way table of means corrected for cultivar and site main effects, $\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$) is modeled directly as a function of the cultivar and environmental variables. A factorial regression model for the mean of the ith cultivar in the jth environment, for which the interaction includes the cultivar covariables x_{i1} to x_{iG}, is

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \sum_{g=1,\,G} x_{ig}\,\xi_{jg} \qquad [5]$$

The parameters μ, τ_i, and β_j are the same as in Eq. [1] and [3]. The GEI consists of the products of the environmental factors ξ_{j1} to ξ_{jG} with respect to cultivar covariables x_{i1} to x_{iG} ($G \leq I - 1$). The cultivar covariables are known, but the environmental potentialities have to be estimated. In matrix notation, Model [5] can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \mathbf{X}\boldsymbol{\Xi}' \qquad [6]$$

where the first three terms on the right side are the same as before, $\mathbf{X} = (x_{ig})$ is a $I \times G$ matrix of known cultivar covariables and $\boldsymbol{\Xi} = (\xi_{jg})$, a $J \times G$ matrix of unknown environmental constants, and G is the number of cultivar covariables.

An FR model in which the interaction part contains the environmental covariables z_{j1} to z_{jH} can be written as:

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \sum_{h=1,\,H} \zeta_{ih}\,z_{jh} \qquad [7]$$

The GEI term in this model allows the cultivars to have different sensitivities, ζ_{i1} to ζ_{iH} ($H \leq J - 1$), to the environmental covariables. The values of the environmental variables are known, but the cultivar sensitivities need to be estimated. Similar to Model [5], Model [7] can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\,\boldsymbol{\beta}' + \boldsymbol{\zeta}\,\mathbf{Z}' \qquad [8]$$

where $\mathbf{Z} = (z_{jh})$ is the $J \times H$ matrix of environmental covariables and $\boldsymbol{\zeta} = (\zeta_{ih})$ is a $I \times H$ matrix of cultivar sensitivities, and H is the number of environmental covariables.

The structure of the FR model including both cultivar and environmental covariables simultaneously is similar to that of Models [5] and [7] (Denis, 1988; van Eeuwijk et al., 1996). In matrix notation it can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \mathbf{X}\boldsymbol{\nu}\mathbf{Z}' + \mathbf{X}\,\boldsymbol{\Xi}' + \boldsymbol{\zeta}\mathbf{Z}', \qquad [9]$$

where in the new term $\mathbf{X}\boldsymbol{\nu}\mathbf{Z}'$, $\boldsymbol{\nu}$ is a $G \times H$ matrix of regression coefficients to cross-products of cultivar and environmental covariables. General identification constraints for factorial regressions with already centered covariables are $\boldsymbol{\zeta}'\mathbf{X} = \mathbf{Z}'\boldsymbol{\Xi} = \mathbf{0}$, where $\mathbf{0}$ is now a matrix of zeros of order $H \times G$. Covariables may be quantitative and qualitative, and more complicated FR models are possible by combining quantitative and qualitative covariables.

In this study the FR procedure was implemented in GENSTAT version 5, release 3.2 (GENSTAT, 1995). The stepwise procedure implemented in Genstat for the multiple linear regression, selects a term to be included or excluded from the model based on an F-test. For example, for X_1, X_2, X_3, and X_4, explanatory variables, the procedure starts by fitting a model containing variable X_1. Then it attempts to drop X_1 and to add, one at the time, X_2, X_3, and X_4. The procedure permanently modifies the current model according to the change that was most successful; if dropping X_1 improves the model, then X_1 is permanently removed; or, when no removals are worthwhile, if adding X_2 gives the biggest improvement, then X_2 is permanently included. The stepwise procedure allows for forward selection or backward elimination.

Partial Least Squares Regression. In many situations, when the number of variables (S) is much larger than the number of observations (N), and there is high collinearity among variables, the usual methods for fitting regressions based on ordinary least squares are not adequate. In this situation, partial least squares regression seems to be a more appropriate alternative. Details of PLS theory (Helland, 1988) and its similarities to principal components regression and stepwise multiple linear regression are described in Aastveit and Martens (1986). A description of univariate and multivariate PLS and their algorithms was given in Vargas et al. (1998). In this paper, the multivariate PLS algorithm, the cross validation procedure, and the F-test were applied by a procedure implemented in GENSTAT version 5, release 3.2 (GENSTAT, 1995).

For the standard situation where multivariate PLS is used to model cultivar responses ($\mathbf{Y}$) over environments on environmental covariables ($\mathbf{Z}$), the corresponding bilinear forms are $\mathbf{Z} = \mathbf{TP}' + \mathbf{E}$ and $\mathbf{Y} = \mathbf{TQ}' + \mathbf{F}$, respectively, where matrix $\mathbf{T}$ contains the Z-scores, matrix $\mathbf{P}$ contains the Z-loadings, matrix $\mathbf{Q}$ contains the Y-loadings, and $\mathbf{E}$ and $\mathbf{F}$ are the residual matrices. It is easiest to work with the transpose of $\mathbf{Y}$: $\mathbf{Y}'$ such that the columns of $\mathbf{Y}'$ (i.e., the rows of $\mathbf{Y}$) contain cultivar responses over environments. Then $E(\mathbf{Y}') = (\mathbf{TQ}')' = \mathbf{QW}'\mathbf{Z}' = \boldsymbol{\zeta}\mathbf{Z}'$, where $\mathbf{T}$ contains the Z-scores (indexed by environments), $\mathbf{W}$ the Z-loadings (or weights, indexed by environmental variables), and $\mathbf{Q}$ the Y-loadings (indexed by genotypes). $\boldsymbol{\zeta}$ contains the PLS approximation to the regression coefficients of the responses in $\mathbf{Y}'$ (genotypic responses) to the explanatory variables in $\mathbf{Z}$ (environmental variables). Note that $\boldsymbol{\zeta}\mathbf{Z}'$ is the same as the last term of Eq. [8]. From this formulation, it can be deduced which biplots can be constructed to summarize PLS analyses. The set $\mathbf{T}$, $\mathbf{W}$, and $\mathbf{Q}$ can be depicted in the same biplot: the rows of matrix $\mathbf{T}$ contain the coordinates for environments, the rows of $\mathbf{W}$ contain the coordinates for environmental covariables and the rows of $\mathbf{Q}$ contain the coordinates for cultivars. Projection of the jth row of $\mathbf{T}$ on the ith row of $\mathbf{Q}$ (or vice versa) approximates the interaction of the ith genotype on the jth environment: $\mathbf{Y}' = (\mathbf{TQ}')'$. Projection of the hth row of $\mathbf{W}$ on the ith row of $\mathbf{Q}$ (or vice versa) approximates the regression coefficient of the ith genotype on the hth environmental covariable: $\mathbf{QW}' = \boldsymbol{\zeta}$. Thus, the PLS biplot including representations of cultivars, environments and covariables allows the same types of interpretation to be made as the enriched AMMI biplot introduced earlier.

When environmental responses over cultivars are modeled on cultivar covariables, $E(\mathbf{Y}) = \mathbf{TQ}' = \mathbf{XWQ}' = \mathbf{X}\boldsymbol{\Xi}'$ (the same as the last term of Eq. [6]) the rows of $\mathbf{T}$ will contain coordinates for the cultivars, the rows of $\mathbf{W}$ will contain coordi-

nates for the cultivar covariates, and the rows of **Q** will contain coordinates for the environments. Ξ' contains the PLS approximation to the regression coefficients of the responses in **Y** to the explanatory variables in **X**. Biplot relations follow from $\mathbf{Y} = \mathbf{TQ}'$ and $\mathbf{WQ}' = \Xi'$.

Experimental Data

Durum Wheat Variety Trial (Data Set 1). This data set, used by Vargas et al. (1998), consisted of one experiment with seven durum wheat cultivars tested for 6 yr (1990–1995) in Ciudad Obregón, Mexico. The cultivars included were a historical set released from the early 1960s to the late 1980s; the order of Numbers 1 to 7 is the order of cultivar releases over time (Sayre et al., 1997).

The cultivar variables, ***X***, were days to anthesis after emergence (ANT), days to maturity after emergence (MAT), days of grainfill (GFI), plant height (cm) (PLH), above-ground biomass (kg ha^{-1}) (BIO), harvest index (HID), straw yield (kg ha^{-1}) (STW), number of spikes per square meter (NSM), number of grains per square meter (NGM), number of grains per spike (NGS), thousand-kernel weight (g) (TKW), weight per tiller (g) (WTI), spike grain weight (g) (SGW), vegetative growth rate (kg ha^{-1} d^{-1}) (VGR), and individual kernel growth rate (mg kernel^{-1} d^{-1}) (KGR) during the grainfill period. The environmental variables, **Z**, measured from December to March of each year were mean daily maximum temperature (°C) (MT), mean daily minimum temperature (°C) (mT), monthly total precipitation (mm) (PR), and sun hours per day (SH).

Wheat Agronomic Trial (Data Set 2). This data set consisted of one wheat experiment including several treatments for cultural practices, conducted over 10 yr (1988–1997) in Ciudad Obregón, Mexico. Each year the experiment was arranged in a randomized complete block design with three replicates. Treatments were obtained by combining four factors at the following levels: tillage at 2 levels (1 = with deep knife, 2 = without deep knife), summer crop at 2 levels [1 = sesbania, (*Sesbania* spp.) 2 = soybean (*Glycine max* (L.) Merr.)], manure at 2 levels (1 = with chicken manure, 2 = without chicken manure), and nitrogen fertilization rate at 3 levels (1 = 0 kg N ha^{-1}, 2 = 100 kg N ha^{-1} and 3 = 200 kg N ha^{-1}), resulting in 2 × 2 × 2 × 3 = 24 treatments. Therefore, Treatment 1 is [1–1–1–1], Treatment 2 is [2–1–1–1], Treatment 3 is [1–2–1–1], and so on up Treatment 24 [2–2–2–3].

Data matrix **Y** had 24 rows (treatments) and 10 columns (years). Matrix **Y** had grain yield interaction residuals ($\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$). The 27 explanatory variables in the **Z** matrix of size 10 × 27 (years × environmental variables) were mean minimum temperature sheltered (°C) (mT), mean minimum temperature unsheltered (°C) (mTU), mean maximum temperature sheltered (°C) (MT), total monthly precipitation (mm) (PR), mean sun hours per day (SH), and total monthly evaporation (mm) (EV). Environmental variables were measured from December through April of each year. All covariables were centered and standardized prior to analysis. For reasons of consistency with earlier analyses (Vargas et al., 1998), the columns of the **Y** matrix were standardized.

RESULTS

Data Set 1

AMMI Analysis, Biplot and Correlations

The combined analysis of variance across years showed that 66% of variation among the 42 cultivar × year combinations was explained by differences among cultivar means, 22% by differences between year means, and 5% by cultivar × year interaction (Table 1). AMMI analysis of variance indicated that the first multiplivative term was significant ($P < 0.05$) by the F_{GH1} test (Cornelius et al., 1993, 1996; Cornelius, 1993) and the first two multiplicative terms were significant ($P < 0.05$) by the F_1 and F_R tests (Cornelius et al., 1993, 1996; Cornelius, 1993). The first bilinear interaction term of the AMMI analysis of variance accounted for 65% of the cultivar × year interaction sum of squares with 10 degrees of freedom and the second for 15% with 8 degrees of freedom. The first two bilinear terms accounted for 80% of the interaction, indicating that with only 18 degrees of freedom, from the 30 degrees of freedom contained in the analysis of variance cultivar × year interaction, a considerable amount of the GEI was explained (Table 1).

To investigate relationships between additive and multiplicative parameters in the AMMI model and the values of the cultivar and environmental covariables, correlations coefficients were calculated between the cultivar mean grain yields and each of the cultivars covariables. Similarly, the environment means for grain yield were correlated with each of the environmental covariable. The coefficients of determination (R^2) for the regression of the standardized cultivar and environmental covariables on the cultivar and environmental scores of the first two bilinear terms (scores of Axes 1 and 2) were also computed. The cultivar main effect was highly positively correlated with number of grains per square meter (NGM), number of grains per spike (NGS), harvest index (HID), spike grain weight (SGW), and aboveground biomass (BIO), and was highly negatively correlated with individual kernel growth rate (KGR) (Table 2). The R^2 values of the regressions of these cultivar variables on the scores of AMMI Axes 1 and 2 were also high. The environmental main effect was negatively correlated with the environmental variables, minimum temperature in December (mTD), January (mTJ), and February (mTF) and precipitation in February (PRF), and positively correlated with maximum temperature

Table 1. AMMI analysis of variance for Data Sets 1 and 2. Data Set 1 consisted of one experiment with seven durum wheat cultivars tested for 6 yr (1990–1995) at Ciudad Obregón, Mexico. The cultivars included were a historical set released from the early 1960s to the late 1980s. Data Set 2 consisted of one wheat experiment including several treatments for cultural practices, conducted over 10 yr (1988–1997) at Ciudad Obregón, Mexico.

Source	df	Sum of squares (× 10^6)†	Mean squares (× 10^6)†	Prob
		Data set 1		
Cultivar	6	183.740	306.233	0.0001
Year	5	62.624	125.248	0.0001
Cultivar × year	30	14.547	4.849	0.0001
Bilinear term 1	10	9.549	9.549	0.0001
Bilinear term 2	8	2.238	2.797	0.0679
Deviation	12	2.760	2.300	0.1169
Error	72	10.410	1.446	
		Data set 2		
Treatment	23	773.970	336.508	0.0001
Year	9	373.260	414.733	0.0001
Year × treatment	207	279.520	13.503	0.0001
Bilinear term 1	31	151.130	48.751	0.0001
Bilinear term 2	29	39.112	13.486	0.0001
Bilinear term 3	27	36.781	13.622	0.0001
Deviations	120	52.497	4.374	0.0001
Error	460	110.870	2.410	

† Actual values multiplied by this factor to obtain reported values.

in December (MTD) and January (MTJ) and sun hours in January (SHJ) and February (SHF).

The AMMI biplot (Fig. 1a) separates the high yielding years, 1990, 1991, and 1994, from the low yielding years, 1993 and 1995, along the fist axis from left to right. With respect to cultivars, along the horizontal first axis, earlier released cultivars, 1 and 2, are separated from intermediate and later released cultivars, 3, 4, 5, 6, and 7. Cultivars 1 and 2 were positively influenced by environmental conditions in 1990, 1991, and 1994 and negatively influenced by environmental conditions in 1993 and 1995. Cultivars 5 and 6 were favored in 1995 and, to some extent, in 1993, while they were negatively influenced by environmental conditions prevailing in 1990, 1991, 1992, and 1994.

Table 2. Correlations coefficients (r) between cultivar covariables versus cultivar mean grain yield and environmental covariables versus environmental mean grain yield, and the proportion of variation explained in each cultivar and environmental variable by the regression of the cultivar and environmental covariables on the scores of the first two AMMI bilinear terms (R^2) for Data Sets 1 and 2.

Data set 1					
Cultivar†			Environmental‡		
Covariable	Correlation	R^2	Covariable	Correlation	R^2
HID	0.94**	0.91	PRF	−0.66	0.95
NGS	0.96**	0.90	MTM	−0.28	0.87
NGM	0.99**	0.88	mTJ	−0.81*	0.85
SGW	0.90**	0.85	SHM	−0.43	0.78
KGR	−0.82*	0.84	SHF	0.68	0.75
PLH	0.71	0.74	SHJ	0.70	0.75
BIO	0.89**	0.66	MTJ	0.55	0.72
ANT	−0.48	0.61	PRJ	−0.54	0.67
WTI	0.47	0.58	PRM	−0.36	0.64
MAT	−0.45	0.58	SHD	0.45	0.46
TKW	−0.67	0.41	MTD	0.76	0.43
VGR	0.63	0.41	MTF	0.12	0.32
NSM	0.10	0.36	mTF	−0.75	0.30
STW	0.01	0.18	mTD	−0.60	0.29
GFI	0.14	0.10	PRD	−0.34	0.08
			mTM	−0.31	0.04

Data set 2					
Environmental‡			Environmental‡		
Covariable	Correlation	R^2	Covariable	Correlation	R^2
EVA	0.17	0.68	MTA	−0.15	0.32
mTM	−0.33	0.67	mTD	−0.30	0.28
mTUM	−0.24	0.62	SHD	0.23	0.25
EVD	0.54	0.58	mTJ	−0.20	0.25
MTF	−0.01	0.53	mTUD	−0.20	0.25
EVJ	0.42	0.52	MTD	0.59*	0.24
SHJ	0.28	0.51	MTJ	0.31	0.24
mTUF	−0.34	0.49	PRM	−0.29	0.22
mTF	−0.41	0.48	mTUA	0.27	0.14
EVF	0.47	0.45	PRD	−0.50	0.14
EVM	0.19	0.36	mTA	0.43	0.08
PRF	−0.37	0.35	SHF	−0.30	0.20
PRJ	−0.25	0.33	MTM	−0.13	0.00
mTUJ	−0.03	0.32			

*, ** Significantly different from zero at the 0.05 and 0.01 levels of probability, respectively.

† HID = harvest index, NGS = number of grains per spike, NGM = number of grains per square meter, SGW = spike grain weight, KGR = individual kernel growth rate, PLH = plant height, BIO = above-ground biomass, ANT = days to anthesis after emergence, WTI = weight per tiller, MAT = days to maturity after emergence, TKW = thousand kernel weight, VGR = vegetative growth rate, NSM = number of spikes per square meter, STW = straw yield, GFI = days for grain filling.

‡ PR = total monthly precipitation, MT = mean maximum temperature sheltered, mT = mean minimum temperature, SH = sun hours per day; EV = total monthly evaporation; mTU = mean minimum temperature unsheltered; D = December, J = January, F = February, M = March, A = April.

To help interpret the AMMI biplot patterns, the directions of greatest changes for six cultivar covariables ($R^2 > 0.73$, Table 2), as obtained from regressions of the standardized covariables on the first and second AMMI axes, were added to the biplot. These were, in decreasing order with respect to R^2, HID, NGS, NGM, SGW, KGR, and PLH. Cultivars 1 and 2 had above average values for KGR (i.e., they were positively associated with covariable KGR) and had below average values for PLH, SGW, NGS, NGM, and HID. The intermediate-released cultivars, 3 and 4, and later released cultivar, 7, had above average values for HID and NGM, and below average values for KGR. Cultivars 5 and 6 had above average values for SGW, PLH, and NGS. Cultivars 1 and 2 had the largest values for KGR; Cultivars 3, 4, and 7 showed the largest values for HID and NGM; and Cultivars 5 and 6 had the largest values for NGS, PLH, and SGW.

The direction of the greatest changes for six environmental covariables ($R^2 > 0.75$) was also added to the AMMI biplot (Fig. 1a). These were in decreasing order with respect to R^2 PRF, maximum temperature in March (MTM), mTJ, sun hours per day in march (SHM), SHF, and SHJ. Minimum temperature in December (R^2 = 0.286) was included in the AMMI biplot for reasons that will be explained later. Years 1993 and 1995 had above average values for MTM and SHM but below average values for SHF. Years 1990, 1991, 1992, and 1994 had high SHF, but low SHM.

Apparently SHM and MTM in 1993 and 1995 caused Cultivars 5 and 6 to develop relatively more NGS and to have heavier SGW than the other cultivars. Years 1990, 1991, 1992, and 1994 had lower values of SHM and MTM and, Cultivars 5 and 6 had correspondingly lower NGS and SGW in those years. Sun hours in January and February (SHJ and SHF) in 1990, 1991, 1992, and 1994 helped Cultivars 1 and 2 develop high individual KGR; however, low SHJ and SHF values in 1993 and 1995 were not conducive for Cultivars 1 and 2 to develop this trait.

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Cultivar Explanatory Variables

Results from the PLS procedure showed that the first and second factors explained 56 and 13% of the cultivar × year interaction, respectively. Table 3 shows the X-loadings (weights) for each cultivar covariable sorted by the first PLS factor, as well as the complete set of individual factorial regressions on each cultivar covariable, ranked by their contribution to the total cultivar × year sum of squares.

With respect to the FR analysis, there were 15 cultivar covariables available, but only a maximum of $G \leq I - 1$ of them can be used simultaneously, where G is the number of cultivar covariables and I is the number of cultivars (7). All the cultivar covariables were included in the individual FR models and those that explained most of the cultivar × year interaction sum of squares were the same as those that had the highest R^2 values with AMMI scores (Table 2), namely, NGS, NGM, SGW, HID, PLH, and KGR. The rank order of the

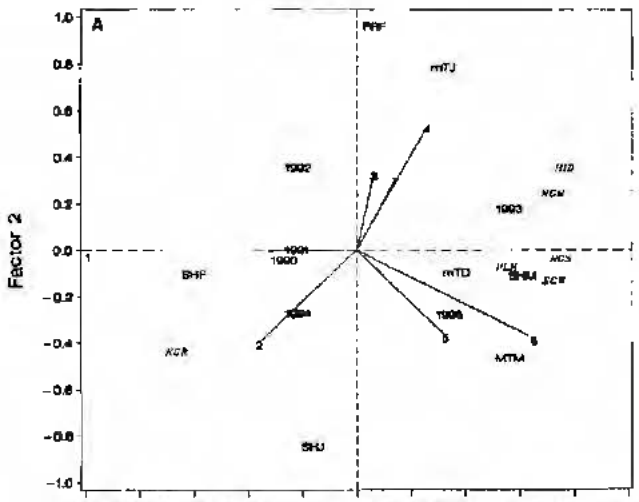

Fig. 1 (A). Biplot of the first and second AMMI axes representing the environmental and cultivar scores of 6 yr (1990–1995) and seven durum wheat cultivars (1–7) enriched with the direction of greatest change of selected cultivar and environmental covariables from Data Set 1. Scaling constant c = 0.5. Cultivar covariables are PLH = plant height, KGR = individual kernel growth rate, HID = harvest index, NGS = number of grains per spike, NGM = number of grains per square meter, SGW = spike grain weight. Environmental covariables are SH = sun hours per day, MT = mean maximum temperature; D = December, J = January, F = February, M = March.

cultivar covariables in relation to how much they contributed to explaining the cultivar × year interaction was practically identical for the PLS approach and the FR model.

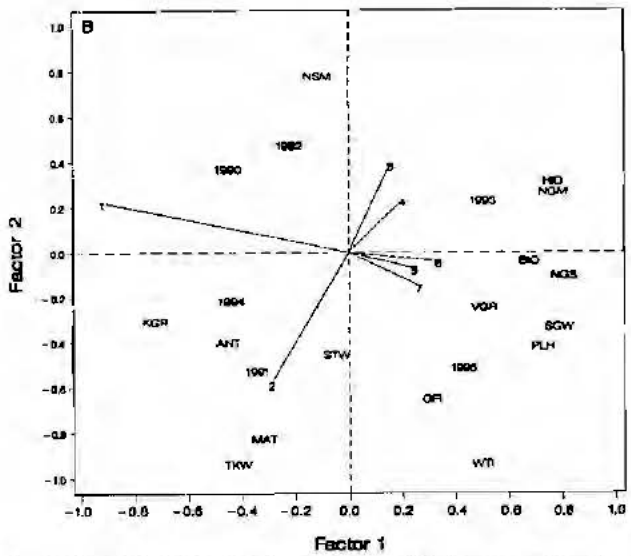

Fig. 1 (B). Biplot of the first and second PLS factors representing the X-scores of seven durum wheat cultivars (1–7), the Y-loadings of 6 yr (1990–1995) enriched with the X-loadings of 15 cultivar covariables from Data Set 1. Scaling constant c = 0.5. Cultivar variables are NGS = number of grains per spike, HID = harvest index, SGW = spike grain weight, NGM = number of grains per square meter, KGR = individual kernel growth rate, PLH = plant height, BIO = above-ground biomass, VGR = vegetative growth rate, ANT = days to anthesis after emergence, WTI = weight per tiller, TKW = thousand kernel weight, MAT = days to maturity after emergence, GFI = days for grain filling, STW = straw yield, NSM = number of spikes per square meter.

The PLS biplot of environmental responses over cultivars versus cultivar covariables is depicted in Fig. 1b. Similarities to the AMMI biplot (Fig. 1a) are evident. The PLS biplot shows that subsets of correlated cultivar covariables can be distinguished (the angle between the variables is important): (HID, NGM), (BIO, NGS, VGR, SGW, PLH), (GFI, WTI), (STW, MAT, TKW, ANT, KGR), and NSM. In contrast to the AMMI biplot, the PLS biplot separated Cultivar 7 from Cultivars 3 and 4, and grouped it with the Cultivars 5 and 6. Cultivar 1 had high KGR and ANT and low HID, NGM, BIO, NGS, VGR, SGW, and PLH. It yielded relatively well in 1990, 1991, 1992, and 1994 and yielded poorly in 1993 and 1995. Cultivars 5, 6, and 7 behaved exactly the opposite. Cultivar 2 had high KGR, ANT, MAT, TKW, and STW, while being low for NSM, HID, and NGM. It yielded relatively well in 1991, 1994, and 1995 and yielded poorly in 1990, 1992, and 1993. Performance of Cultivars 3 and 4 was the reverse of Cultivar 2 with respect to years.

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Environmental Explanatory Variables

The first PLS factor explained 40% of the cultivar × year interaction sum of squares, while the second PLS factor explained 26%. Table 3 shows the Z-loadings (weights) of the environmental explanatory variables sorted by the first PLS factor, and the individual factorial regression for each of the environmental variables, ranked by their contribution to the analysis of variance GEI sum of squares. The maximum number of covariables that can be used simultaneously in factorial regressions with centered environmental variables is $H \leq J -$

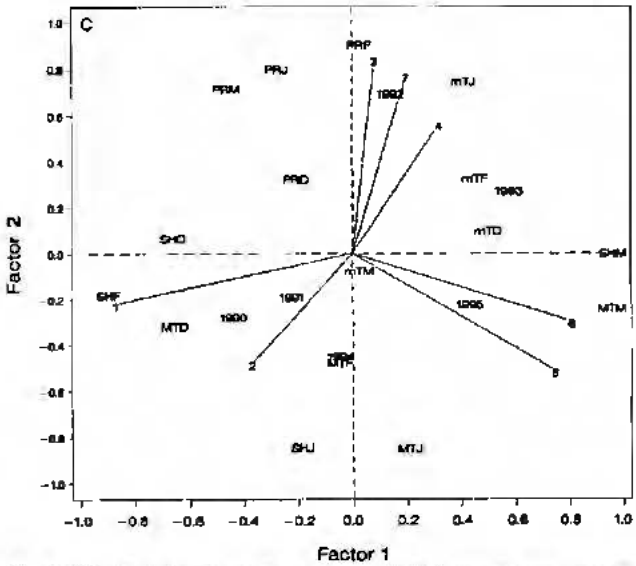

Fig. 1 (C). Biplot of the first and second PLS factors representing the Z-scores of 6 yr (1990–1995) and the Y-loadings of seven durum wheat cultivars (1–7) enriched with the Z-loadings of 16 environmental covariables from Data Set 1. Scaling constant c = 0.5. Environmental variables are mT = mean minimum temperature, MT = mean maximum temperature, PR = total monthly precipitation, SH = sun hours per day; D = December, J = January, F = February, M = March.

1 (see Eq. [7]). As $J = 6$ yr in this data set, $H \leq 5$. The environmental variables that explained most of the cultivar × year interaction sum of squares were SHM, SHF, MTM, MTD, and SHD (Table 3). SHM, SHF, and MTM had values for $R^2 > 0.75$ (Table 2).

The PLS method determined that environmental covariables SHM, SHF, SHD, MTM, and MTD were associated with Factor 1, which explained a large proportion of the cultivar × year interaction (i.e., they had the highest absolute loadings) (Table 3). The FR method also considered these covariables individually to be the five most important environmental covariables in explaining cultivar × year interaction. All other environmental covariables were ranked similarly by both methods. Both procedures considered precipitation to be less important for explaining cultivar × year interaction.

The PLS biplot depicted in Fig. 1c is similar to the AMMI biplot (Fig. 1a) enriched with the directions of greatest change for the environmental covariables with $R^2 > 0.75$. Unlike the AMMI biplot, the PLS biplot separated the low-yielding year 1992 from the other two low yielding years 1994 and 1995. Years 1990 and 1991 had high SHF and MTD and low mTJ and mTF. In contrast to 1990 and 1991, 1993 had high mTJ and mTF, and low MTD and SHF. The year 1992 had high precipitation in general (PRD, PRJ, PRF, PRM) and high mTJ, while 1992 had low MTF, SHJ, and MTJ. In contrast to 1992, 1994 had high MTF, SHJ and MTJ, and low precipitation, low mTJ and mTF. Precipitation during the months of the growing cycle (PRD, PRJ, PRF, and PRM) formed a subset of correlated covariables located in the upper left quadrant of the PLS biplot, whereas a subset of minimum temperatures during the growing cycle (mTD, mTJ, and mTF) is located in the upper right quadrant (mTM is at the center of the PLS biplot). Variables SHM and MTM were positively correlated and formed a subset of environmental variables with high loadings for the first PLS factor.

Cultivars 1 and 2 were favored by SHJ, SHF, MTD, and MTF. This led to higher yields in 1990, 1991, and 1994. Lower mTD, mTJ, and mTF and greater PRF did not favor Cultivars 1 and 2, most notably in 1992 and 1993; however, these environmental conditions during the 1992 and 1993 growing cycles favored Cultivars 3,

Table 3. X-loadings and Z-loadings of the cultivar and environmental covariables, respectively, of the first two PLS factors (sorted by the first PLS factor), and mean squares of all individual factorial regressions for Data Set 1.

Covariable	Partial least squares		Factorial regression			
	Factor 1†	Factor 2	Source	df	Mean square (× 10^5)	Prob > F
	X-loadings		Cultivar covariables‡			
			Year × Cultivar	30	4.849	0.0001
NGS	0.36	0.04	NGS	5	17.298	<0.0001
SGW	0.35	0.14	NGM	5	16.349	<0.0001
NGM	0.34	−0.14	SGW	5	16.339	<0.0001
HID	0.34	−0.14	HID	5	16.058	<0.0001
PLH	0.32	0.18	PLH	5	14.469	<0.0001
KGR	−0.32	0.13	KGR	5	14.290	<0.0001
BIO	0.30	0.01	BIO	5	12.825	<0.0001
VGR	0.23	0.11	VGR	5	8.574	0.0001
WTI	0.22	0.41	WTI	5	8.156	0.0001
ANT	−0.20	0.17	TKW	5	7.291	0.0005
TKW	−0.18	0.41	ANT	5	7.108	0.0006
MAT	−0.14	0.36	MAT	5	5.466	0.0042
GFI	0.14	0.29	GFI	5	3.972	0.0251
NSM	−0.05	−0.50	NSM	5	3.367	0.0512
STW	−0.02	0.20	STW	5	1.563	0.3785
			Error	72	1.446	
			Environmental covariables¶			
	Z-loadings		Year × Cultivar	30	4.849	0.0001
SHM	0.47	0.01	SHM	6	12.594	<0.0001
MTM	0.44	−0.11	SHF	6	12.112	<0.0001
SHF	−0.42	−0.08	MTM	6	11.540	<0.0001
SHD	−0.31	0.04	MTD	6	7.745	0.0001
MTD	−0.30	−0.14	SHD	6	7.086	0.0003
mTD	0.23	0.05	mTJ	6	5.815	0.0016
PRM	−0.22	0.34	mTF	6	5.786	0.0017
mTF	0.21	0.16	mTD	6	5.484	0.0024
mTJ	0.19	0.35	PRM	6	4.582	0.0082
PRJ	−0.13	0.38	SHJ	6	3.850	0.0218
PRD	−0.10	0.15	PRF	6	3.748	0.0249
MTJ	0.10	−0.39	PRJ	6	3.446	0.0373
SHJ	−0.09	−0.39	MTJ	6	3.156	0.0547
MTF	−0.02	−0.22	MTF	6	2.616	0.1111
mTM	0.01	−0.03	mTM	6	2.362	0.1512
PRF	0.01	0.42	PRD	6	1.805	0.2962
			Error	72	1.446	

† PLS results extracted from Tables 1 and 2 of Vargas et al. (1998).

‡ NGS = number of grains per spike, SGW = spike grain weight, NGM = number of grains per square meter, HID = harvest index, PLH = plant height, KGR = individual kernel growth rate, BIO = above-ground biomass, VGR = vegetative growth rate, WTI = weight per tiller, ANT = days to anthesis after emergence, TKW = thousand kernel weight, MAT = days to maturity after emergence, GFI = days for grain filling, NSM = number of spikes per square meter, STW = straw yield.

¶ SH = sun hours per day, MT = mean maximum temperature, mT = mean minimum temperature, PR = total monthly precipitation; D = December, J = January, F = February, M = March.

Table 4. Analysis of variance tables for stepwise multiple factorial regression models with environmental and cultivar covariables, for Data Sets 1 and 2. Terms in factorial regression models appear in the order of inclusion.

Source	d.f.	Sum of squares ($\times 10^6$)	Mean squares ($\times 10^5$)	F	Prob > F
			Data set 1		
			MFR1†		
Cultivar × SHM‡	6	7.557	12.595	8.71	<0.0001
Cultivar × mTD	6	2.655	4.425	3.06	0.0100
Cultivar × PRF	6	2.350	3.916	2.70	0.0198
Deviations	12	1.986	1.655	1.14	0.3428
			MFR2		
Year × NGS	5	8.649	17.298	11.96	<0.0001
Year × NGM	5	1.782	3.564	2.46	0.0409
Deviations	20	4.115	2.057	1.42	0.1489
			MFR3		
NGS × SHM	1	6.484	64.840	44.84	<0.0001
NGS × mTD	1	1.947	19.470	13.46	0.0004
NGS × PRF	1	0.748	7.480	5.17	0.0259
Deviations	27	5.364	1.986	1.37	0.1464
			Data set 2		
Treat × mTF¶	23	78.53	34.143	14.16	0.0001
Treat × EVF	23	54.75	23.804	9.87	0.0001
Treat × mTJ	23	40.50	17.608	7.30	0.0001
Treat × mTUM	23	27.62	12.008	4.98	0.0001
Deviations	115	78.12	6.793	2.81	0.0001

† MFR1 = Multiple Factorial Regression model with environmental covariables; MFR2 = Multiple Factorial Regression model with cultivar covariables; MFR3 = Multiple factorial regression model with cross products of cultivar and environmental covariables from MFR1 and MFR2.
‡ SHM = sun hours per day in March; mTD = minimum temperature in December; PRF = precipitation in February; NGS = number of grains per spike; NGM = number of grains per square meter.
¶ mTF = minimum temperature sheltered in February; EVF = evaporation in February; mTJ = minimum temperature sheltered in January; mTUM = minimum temperature unsheltered in March.

4, and 7. Cultivars 5 and 6 had high yields in year 1995, probably because of higher MTJ, MTM, and SHM. Covariables SHM, MTM and SHF had high loading values for the first PLS factor, whereas covariables PRF, MTJ, SHJ, and PRJ had high loading values for the second PLS factor. These seven covariables had the highest R^2 values (Table 2).

Explaining Genotype × Environment Interaction Using Multiple Factorial Regression with Cultivar and Environmental Explanatory Variables Simultaneously

Multiple factorial regression coupled with a stepwise procedure for variable selection was used to search for informative sets of environmental and/or cultivar covariables. When the set of independent variables from which a selection had to be made consisted of all environmental covariables, a model for the interaction was found that included the terms cultivar × SHM, cultivar × mTD, and cultivar × PRF (Table 4). This model explained 86% of the GEI with 18 df. This model, called multiple factorial regression 1 (MFR1), appeared be slightly better than the $AMMI_2$ (with two bilinear terms) model, which explained 81% of the GEI with the same 18 df (Table 1).

When the set of candidate variables consisted of all cultivar variables, a model was found that included the terms Year × NGS and Year × NGM (Table 4). This model, MFR2, explained 72% of the GEI with 10 df and appeared to be superior to $AMMI_1$ (with only one bilinear term), which accounted for 66% of the GEI with 10 df (Table 1). Thus, AMMI models with one or two bilinear terms were possibly less effective than the MFR1 and MFR2 models. It should be pointed out that although MFR1 and MFR2 represent the best sets of environmental and cultivar covariables that were found by stepwise regression, a number of other sets were equally good.

Relating significant cultivar and environmental covariables obtained in the stepwise multiple factorial regression (Table 4) to the clusters of environmental and cultivar covariables previously described in the PLS biplots (Fig. 1b and 1c) can be informative. For example, in Fig. 1c, environmental covariables, SHM and MTM, formed a cluster, and the stepwise procedure selected SHM as the more important. Covariables, mTD, mTJ, mTF, and mTM, formed another cluster, and the stepwise procedure selected mTD as a representative candidate for describing GEI. All precipitation covariables formed another cluster, and stepwise regression found PRF to be the best candidate among them. Although mTD and PRF were not among the best covariables for explaining GEI when performing individual FR, they were important when considered together with other environmental covariables. The stepwise variable selection procedure with factorial regression selected, in order, SHM, mTD, and finally PRF. In the PLS biplot (Fig. 1b), roughly four clusters of cultivar covariables may be distinguished, one for each quadrant. The stepwise procedure selected, as significant contributors to explaining GEI, first NGS and then NGM.

After having found 'best' multiple factorial regression (MFR) models for cultivar and environmental variables, we investigated whether further parsimony could be achieved by fitting a multiple FR model that included compound covariables consisting of the products of cultivar and environmental covariables (i.e., we fit a multi-

ple FR model in which only the term $\mathbf{X}\nu\mathbf{Z}'$ from Eq. [9] is maintained). A multiple FR model (MFR3) including the cross products of the cultivar covariable NGS with the environmental covariables SHM, mTD, and PRF gave a very good fit and, on the basis of the differences in sums of squares and df, was not significantly different from MFR1. This gave a very efficient description explaining 63% of the interaction with only 3 df. It is worthwhile to take a look at the interpretation of these cross products. The most important term, NGS × SHM, explained 45% of the total cultivar × year interaction with 1 df. The sign of the estimated coefficient for this term was positive. Thus, cultivars with above average NGS (covariables were all centered, so that positive values after centering mean above average values, and negative values mean below average values) did relatively well in years with above average SHM [(+ NGS in $\mathbf{X}$) × (+ coefficient in ν) × (+ SHM in $\mathbf{Z}'$) = +], as did cultivars with below average NGS in years with below average SHM [(− NGS in $\mathbf{X}$) × (+ coefficient in ν) × (− SHM in $\mathbf{Z}'$) = +). High NGS in years with low SHM and low NGS in years with high SHM are associated with relatively poor performance [(+ NGS in $\mathbf{X}$) × (+ coefficient in ν) × (− SHM in $\mathbf{Z}'$) = −].

The interactions due to NGS × mTD (positive coefficient) and NGS × PRF (negative coefficient) can be interpreted in the same way, although it should be noted that these terms were far less important than the NGS × SHM term.

Data Set 2

AMMI Analysis, Biplot and Correlations

The main effect of treatments (cultural practices) explained 50% of the total sum of squares, whereas differences among year means contributed 24% and the interaction term, 18% (Table 1). The F_R and F_{GH1} tests (Cornelius et al., 1993, 1996; Cornelius, 1993) indicated that the first 5 multiplicative terms were significant ($P < 0.05$) (the first six multiplicative terms were significant by the F_1 test). The first bilinear interaction term of the AMMI model accounted for 54% of the GEI sum of squares, the second 14%, and the third 13%, using 31, 29, and 27 df, respectively. The first two bilinear terms used 60 df of the total of 207 available in the interaction.

Year main effect was not highly correlated with any environmental variable, except maximum temperature sheltered in December (MTD, $r = 0.59$) (Table 2). Total monthly evaporation and mean precipitation in December (EVD and PRD, respectively) showed relatively high correlations with environmental main effects ($r = 0.54$ and -0.50, respectively), indicating the influence of the prevailing climatic conditions on grain yield. In general, values of R^2 obtained from the regression of the standardized environmental variables on the first two bilinear factor scores were relatively low. Only seven variables out of 27 had $R^2 > 0.50$.

The AMMI biplot (Fig. 2a) shows that the axis for the first bilinear term separated the four of the highest yielding years (1994, 1988, 1997, and 1993) from the four lowest yielding years (1995, 1992, 1989, and 1996), althougth 1991, the second highest yielding year, was located with the lowest yielding years; and 1990, the eighth highest yielding year, was grouped with the highest yielding years. Regarding cultural practices, the first axis separated the nine highest yielding treatments (9 [1–1–1–2], 19 [1–2–1–3], 21 [1–1–2–3], 17 [1–1–1–3], 11 [1–2–1–2], 12 [2–2–1–2], 10 [2–1–1–2], 23 [1–2–2–3], and 18 [2–1–1–3]) (five treatments had 200 kg N ha^{-1} and four had 100 kg N ha^{-1}) from the nine treatments with the lowest grain yield (1 [1–1–1–1], 2 [2–1–1–1], 3 [1–2–1–1], 4 [2–2–1–1], 5 [1–1–2–1], 6 [2–1–2–1], 7 [1–2–2–1], 8 [2–2–2–1], and 16 [2–2–2–2]). All had 0 kg N ha^{-1}, except Treatment 16 which had 100 kg N ha^{-1}. The remaining treatments did not show any apparent pattern. The highest yielding treatments were positively related to the highest yielding years, while the lowest yielding treatments were associated with the lowest yielding years.

The AMMI biplot was enriched with the directions of greatest changes for the seven environmental covariables with $R^2 > 0.50$. These covariables were total monthly evaporation in December, January, and April (EVD, EVJ, and EVA, respectively), mean minimum temperature sheltered and unsheltered in March (mTM and mTUM, respectively), mean maximum temperature in February (MTF), and sun hours per day in January (SHJ) (Table 2). Years, 1988, 1990, 1991, and 1996, had above average values (i.e., were positively associated with the covariables EVD, EVJ, EVA, SHJ, and MTF) and had below average values for mTM and mTUM. Years, 1989, 1992, 1994, and 1995, had above average values for covariables, mTM and mTUM, and below average values for the other environmental covariables (Fig. 2a).

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Environmental Explanatory Variables

The cross validation assessment and Osten's (1988) F-test for the number of significant PLS factors indicated that the first factor was significant for prediction, explaining 19% of the year × treatment interaction (data not shown). The second factor explained 28% of the interaction and was found not significant. However, the cross validation gave a PRESS (Predicted Residual Sum of Squares) that was lower than that obtained for the first factor indicating that the second PLS factor improved the prediction accuracy of the model. The first PLS factor had relatively high negative Z-loadings for environmental variables EVD, EVJ, EVF, EVM, and MTD (Table 5) and relatively high positive Z-loadings for mTUF, mTF, mTD, mTM, MTA and PRF. The second PLS factor had high negative loadings for the covariables MTF, mTF and MTA and positive loadings for mTUJ and mTJ.

The maximum number of covariables that could have been used simultaneously in the FR analysis was $H \leq J - 1$ (see eq. [7]), where $J = 10$ (years) so that $H \leq 9$. Although all individual factorial regressions for the

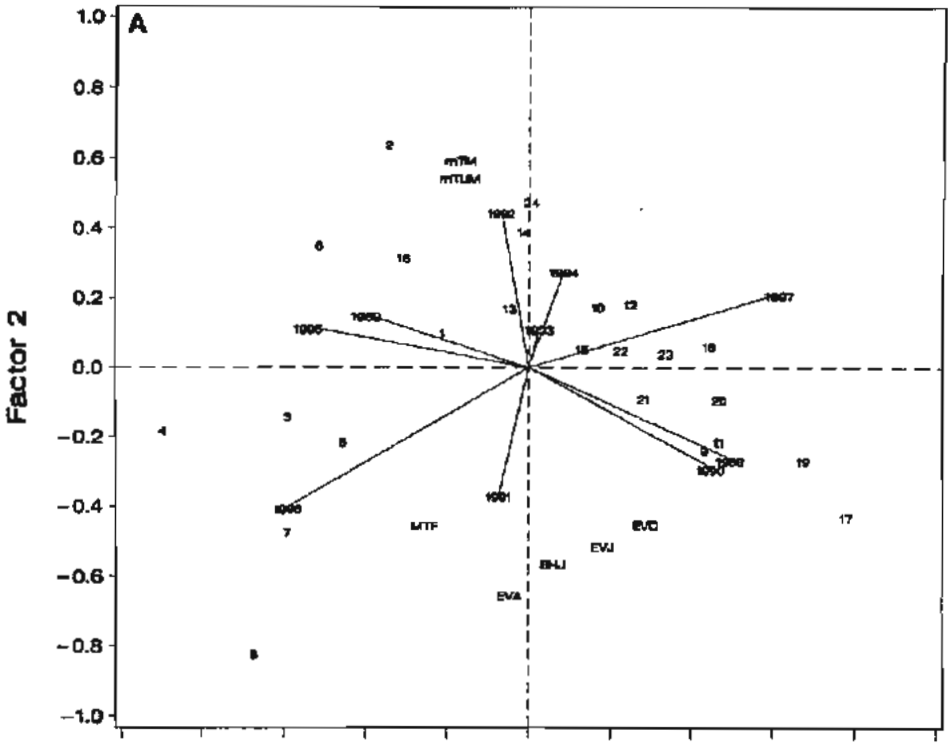

Fig. 2. (A) Biplot of the first and second AMMI axes representing the environmental and treatments scores for 10 yr (1988–1997) and 24 cultural practice treatments (1–24) enriched with the direction of greatest changes of selected environmental covariables from Data Set 2. Scaling constant c = 0.5. Environmental covariables are mT = mean minimum temperature sheltered, mTU = mean minimum temperature unsheltered, EV = total monthly evaporation, MT = mean maximum temperature; D = December, J = January, F = February, M = March, A = April.

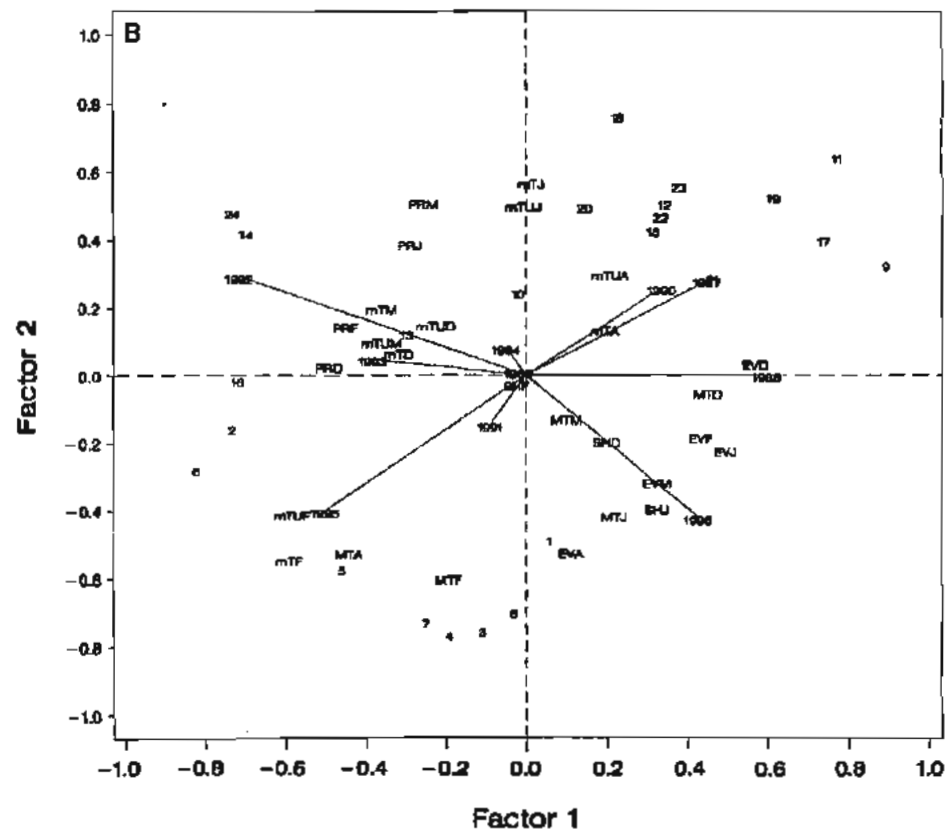

Fig. 2. (B) Biplot of the first and second PLS factors representing the Z-scores of the 10 yr (1988–1997), and the Y-loadings of the 24 cultural practice treatments (1–24) enriched with the Z-loadings of 27 environmental variables from Data Set 2. Environmental variables are mT = mean minimum temperature sheltered, mTU = mean minimum temperature unsheltered, MT = mean maximum temperature sheltered, SH = sun hours per day, EV = total monthly evaporation, PR = total monthly precipitation; D = December, J = January, F = February, M = March, A = April.

centered environmental covariables were significant (each with 23 df), the most interesting were those with the largest sum of squares. The FR model showed that environmental variables, mTF, mTUF, EVD, MTA, MTF, EVJ, mTUM, mTM, and EVF, were important in explaining year × treatment interaction; these variables also had the highest R^2 values for the addition to the AMMI biplot (Table 2). Evaporation in April (EVA) had the largest R^2 value but ranked 14th in FR and 21th in PLS. The rank order of the environmental variables with respect to their contribution to explaining the year × treatment interaction showed good correspondence between PLS and FR for 23 of the covariables (Table 5) (they are ranked at distances lower than four places apart). The most divergent ranking was for MTF, which ranked fifth by FR and 26th by PLS; however, MTF had the highest Z-loading for the second PLS factor (−0.4452). Other variables that differed markedly in ranking were mTUM, EVA, and SHD.

The PLS biplot (Fig. 2b) showed that, for treatments, the results were similar to those obtained with the AMMI biplot (Fig. 2a). The first two PLS factors clearly separated eight of the nine highest yielding treatments (9, 19, 21, 17, 11, 12, 23, and 18) from the nine lowest yielding treatments (1, 2, 3, 4, 5, 6, 7, 8, and 16) (Fig. 2b); however, the separation of years was not as distinct as it was in the AMMI biplot. The low-yielding treatments, 1, 2, 3, 4, 5, 6, 7, 8, and 16, had positive interaction in years with high mTF and mTUF and with high MTF and MTA. This positive interaction was most noticeable in 1995. The year 1995 can be further characterized as being low in mTJ, mTUA, mTA, EVD, MTD, EVF, and EVJ. Negative interactions occurred for the low-yielding treatments in 1988, 1990, and 1997. These years scored just the opposite on the variables enumerated for 1995. In contrast, the eight highest-yielding treatments did relatively well in 1988, 1990, and 1997 and relatively poorly in 1995.

Explaining Genotype × Environment Interaction Using Multiple Factorial Regression with Environmental Explanatory Variables

At least eight covariables were found to be significant by the stepwise selection procedure. The FR model including mTF, EVF, and mTJ had 69 df and explained 62% of the GEI (Table 4), whereas the $AMMI_2$ (with two bilinear terms) accounted for 68% of the GEI with 60 df (Table 1). The factorial regression model with mTF, EVF, mTJ, and mTUM had 92 df and explained 72% the GEI, whereas $AMMI_3$ accounted for 81% of the interaction using 87 df. For this data set, AMMI with two or three bilinear terms was slightly more efficient in describing GEI than FR with three or four of the most significant environmental covariables; however, PLS analysis and stepwise FR are still useful for investigating the influence of different environmental covariables.

The PLS biplot (Fig. 2b) contains roughly four clusters of environmental covariables (one for each quadrant). For example, the first cluster is in the lower left quadrant of Fig. 2b and includes correlated variables mTF, mTUF, MTA, and MTF (in decreasing order according to the sum of squares in the individual FR). The second cluster is in the lower right quadrant and comprises correlated variables EVD, EVJ, EVF, MTD, EVA, SHJ, EVM, SHD, MTJ, and MTM. The third cluster involves mTUA, mTUJ, mTJ, and mTA and the fourth cluster is composed of mTUM, mTM, mTD,

Table 5. Z-loadings of environmental variables sorted by the first PLS factor and mean squares of all individual factorial regressions for Data Set 2.

Environmental covariable	Partial least squares		Factorial regression			
	Factor 1	Factor 2	Source	df	Mean square (× 10^4)	Prob >F
	Z-loadings		Year × Treat	207	1.350	0.0001
EVD†	−0.33	0.05	mTF	23	3.414	<0.0001
EVJ	−0.28	−0.07	mTUF	23	3.122	<0.0001
mTUF	0.27	−0.28	EVD	23	2.634	<0.0001
EVF	−0.27	−0.07	MTA	23	2.522	<0.0001
mTF	0.26	−0.35	MTF	23	2.182	<0.0001
MTA	0.24	−0.31	EVJ	23	1.813	<0.0001
mTD	0.24	0.01	mTUM	23	1.729	<0.0001
mTM	0.23	0.04	mTM	23	1.685	<0.0001
MTD	−0.23	0.01	EVF	23	1.633	<0.0001
PRF	0.22	0.04	mTD	23	1.474	<0.0001
EVM	−0.21	−0.15	PRD	23	1.382	<0.0001
SHJ	−0.20	−0.19	MTD	23	1.342	<0.0001
PRD	0.20	−0.02	PRF	23	1.293	<0.0001
mTUM	0.19	0.02	EVA	23	1.272	<0.0001
mTUD	0.19	0.05	SHJ	23	1.248	<0.0001
SHD	−0.18	−0.10	EVM	23	1.235	<0.0001
PRJ	0.16	0.19	mTUA	23	1.234	<0.0001
PRM	0.14	0.20	mTUD	23	1.091	<0.0001
mTUA	−0.12	0.19	mTUJ	23	1.054	<0.0001
mTA	−0.11	0.10	mTJ	23	1.049	<0.0001
EVA	−0.10	−0.28	PRJ	23	1.034	<0.0001
MTJ	−0.09	−0.19	PRM	23	1.031	<0.0001
mTUJ	0.08	0.29	SHD	23	1.003	<0.0001
MTM	−0.07	−0.07	mTA	23	0.887	<0.0001
mTJ	0.07	0.29	MTJ	23	0.770	<0.0001
MTF	0.03	−0.45	SHF	23	0.610	0.0001
SHF	−0.03	−0.01	MTM	23	0.456	0.0079
			Error	460	0.241	

† EV = total monthly evaporation; mTU = mean minimum temperature unsheltered; mT = mean minimum temperature sheltered; MT = mean maximum temperature sheltered; PR = total monthly precipitation; SH = sun hours per day; D = December; J = January; F = February; M = March; A = April.

PRD, PRF, mTUD, PRJ, PRM, and SHF. It is interesting to note that the stepwise FR selected one covariable from each cluster in the following order: mTF, EVF, mTJ, and mTUM, from the first to the fourth clusters, respectively. The next four covariables selected by the stepwise procedure were from the first (MTA), fourth (PRF and PRJ), and third (mTUA) clusters, respectively. These results indicate that PLS was effective in grouping correlated covariables and that stepwise FR was sensitive enough to detect these groups of correlated covariables and to select the most representative from each cluster.

DISCUSSION

Results of this study indicated that FR and PLS were effective in detecting the environmental and cultivar covariables that explained a sizeable proportion of the total GEI variability in two complex data sets. The AMMI biplot enriched with the covariables that showed high R^2 values was also useful for interpreting GEI of grain yield; however, FR and PLS directly incorporate the external variables into their models, whereas AMMI does not. For Data Set 1, the three procedures identified similar cultivar and environmental covariables that explained most of the GEI. For Data Set 2, results were not as clear as those for Data Set 1, but there was a relatively good correspondence between PLS and FR for 23 of 27 environmental covariables.

In general, the AMMI biplot and the PLS biplot offered similar interpretations of the results for both data sets. The AMMI biplot was very similar to the PLS biplot. Interpretation of these biplots is useful for researchers because it helps to identify major environmental (or cultivar) variables that cause positive or negative interactions between subsets of cultivars with subsets of environments. One advantage of the PLS approach is that a large number of environmental (or cultivar) covariables can be used. Furthermore, PLS is insensitive to multicollinearity; for example, for Data Set 2, minimum and maximum temperatures (sheltered and unsheltered), sun hours per day, and total monthly evaporation are correlated. In the AMMI-enriched biplots, multicollinearity is not a problem but when a large number of genotypic or/and environmental covariables are included, none of them may have sufficiently high R^2 to be drawn in the AMMI biplot. In PLS, the cross validation assessment and Osten's (1988) F-test can be used to test for the significance of the number of components that must be retained. Although the X- or Z-loadings for each covariable for a given PLS factor are not statistical tests of significance, they do provide a measure of their relative importance for explaining GEI.

The main advantage of the FR is that parameters are estimated and hypotheses are tested in relation to the available external covariables. When environmental and cultivar covariables are considered simultaneously, multiple FR with a stepwise variable selection procedure provides a useful tool for selecting the most relevant covariables, and their cross products, for explaining GEI. For both data sets, selected covariables obtained from stepwise FR represented each of the covariable clusters observed in the PLS biplots. While the PLS analysis is done separately on the set of environmental variables and the set of genotypic covariables, FR and the enriched AMMI-biplot perform a simultaneous analysis on both sets of covariables.

When a large number of correlated environmental (and/or cultivar) covariables is available, an important question that researchers face is how to select a set of relevant environmental and cultivar covariables that effectively explain most of the GEI variability. On the basis of the results obtained in this study, a possible strategy for selecting the most important covariables affecting GEI would be to use, first, the PLS analysis with the PLS biplot. It would also be useful to enrich the AMMI biplots with the relevant environmental and cultivar covariables to compare and confirm results obtained by the PLS approach. Results concerning the relevant covariables affecting GEI obtained by PLS and AMMI can always be confirmed by computing factorial regressions. It is therefore advisable to include in the selected subset covariables that are only slightly correlated. An option would be to select the covariables with the largest explained sum of squares in each of the PLS clusters. After arriving at a satisfactory FR model, one could try to reduce further the model by studying just the cross products of the selected environmental and cultivar covariables.

This study indicated that AMMI, PLS, and FR are useful tools for interpreting GEI in the context of multi-environment trials when a large number of external environmental and cultivar covariables are included. The PLS and FR analyses complement each other and offer an aid to researchers not only for determining the importance of individual environmental and cultivar covariables in explaining GEI, but also for finding subsets of covariables that adequately describe GEI in terms of understandable covariables.

ACKNOWLEDGMENTS

The authors thanks the valuable comments and suggestions given by three anonymous reviewers and the associate editor that significantly improved the quality of the manuscript.

REFERENCES

Aastveit, H., and H. Martens. 1986. ANOVA interactions interpreted by partial least squares regression. Biometrics 42:829–844.

Cornelius, P.L. 1993. Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. Crop Sci. 33:1186–1193.

Cornelius, P.L., J. Crossa, and M. Seyedsard. 1996. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. p. 199–234. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction, CRC Press, Boca Raton, FL.

Crossa, J. 1990. Statistical analyses of multilocation trials. Adv. Agron. 44:55–85.

Denis, J.-B. 1988. Two-way analysis using covariates. Statistics 19:123–132.

Denis, J.-B., and J.C. Gower. 1994. Biadditive models. Biometrics 50:310–311.

Eberhart, S.A., and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36–40.

Finlay, K.W., and G.N. Wilkinson. 1963. The analysis of adaptation in a plant breeding programme. Aust. J. Agric. Res. 14:742–754.

Gabriel, K.R. 1971. Biplot display of multivariate matrices with application to principal component analysis. Biometrika 58:453–467.

Gabriel, K.R. 1978. Least squares approximation of matrices by addi-

tive and multiplicative models. J. Roy. Stat. Soc. Series B. 40: 186–96.
Gauch, H.G., Jr. 1988. Model selection and validation for yield trials with interaction. Biometrics 44:705–715.
GENSTAT. 1995. Genstat 5 release 3.2. reference manual supplement. Clarendon Press. Oxford, UK.
Golloh, H.F. 1968. A statistical model which combines features of factor analysis and analysis of variance techniques. Psychometrika 33:73–115.
Helland, I.S. 1988. On the structure of partial least squares regression. Commun. Statist. Simula. 17(2):581–607.
Kempton, R.A. 1984. The use of biplot in interpreting variety by environment interactions. J. Agric. Sci. (Cambridge) 103:123–135.
Mandel, J. 1971. A new analysis of variance model for non-additive data. Technometrics 13:1–18.
Osten, D.W. 1988. Selection of optimal regression models via cross-validation. J. Chemometrics 2:39–48.
Sayre, K.D., S. Rajaram, and R.A. Fischer. 1997. Yield potential progress in short bread wheats in northwest Mexico. Crop Sci. 37:36–42.
Talbot, M., and A.V. Wheelwright. 1989. The analysis of genotype × environment interactions by partial least squares regression. Biuletyn Oceny Odmian Zeszyt 21–22:19–25.
van Eeuwijk, F.A. 1995. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. Euphytica 84:1–7.
van Eeuwijk, F.A. 1996. Between and beyond additivity and non-additivity; the statistical modelling of genotype by environment interaction in plant breeding. Ph.D. Diss., Wageningen University, Wageningen, the Netherlands.
van Eeuwijk, F.A., J.-B. Denis, and M.S. Kang. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. p. 15–49. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction, CRC Press, Boca Raton, FL.
Vargas, M., J. Crossa, K. Sayre, M. Reynolds, M.E. Ramírez, and M. Talbot. 1998. Interpreting genotype × environment interaction in wheat using partial least squares regression. Crop Sci. 38:679–689.
Yates, F., and W.G. Cochran. 1938. The analysis of groups of experiments. J. Agric. Sci. (Cambridge) 28:556–580.

Minimum Sample Size and Optimal Positioning of Flanking Markers in Marker-Assisted Backcrossing for Transfer of a Target Gene

Matthias Frisch, Martin Bohn, and Albrecht E. Melchinger*

ABSTRACT

In recurrent backcrossing designed for introgression of a target allele from a donor into the genetic background of a recurrent parent (RP), molecular markers can accelerate recovery of the recurrent parent genome (RPG). The objectives of this study were to determine in marker-assisted backcrossing (MAB) (i) the optimum distances (d_1, d_2) between the flanking markers and the target locus and (ii) the minimum number of individuals (n) required for obtaining with a certain probability a given number of individuals that carry the donor allele at the target locus and have a minimum proportion of donor genome on the carrier chromosome. Analytical solutions and tabulated results are given for relevant parameters (d_1, d_2, n) required to obtain, with a specified probability of success, at least one desired individual. They depend on the length of the carrier chromosome, the chromosomal position of the target locus, its distance to the flanking marker loci, and the number of individuals evaluated. Our approach can increase the efficiency of MAB by reducing the number of individuals and marker data points required.

Recurrent backcrossing is a breeding method commonly employed to transfer alleles at one or more loci from a donor to a recurrent parent (Allard, 1960). Examples include the transfer of resistance alleles from a wild or unimproved form into elite breeding materials and cultivars or the transfer of a target allele introduced by genetic transformation into a line that is easy to handle in tissue culture but otherwise of no agronomic value (Ragot et al., 1995). Besides transfer of the target allele(s), the main goal is to recover the RPG as completely and as quickly as possible.

Molecular markers are used in recurrent backcrossing for two purposes: (i) as a diagnostic tool for tracing the presence of a target allele, for which direct selection is difficult or impossible (e.g., recessive alleles expressed at a late stage in plant development or quantitative trait loci) and/or (ii) for identifying individuals with a low proportion of the undesirable genome from the donor parent. Adopting the terminology of Hospital and Charcosset (1997), we refer to the first approach as *foreground selection* (for review see Melchinger, 1990) and to the second approach as *background selection* (for review see Visscher et al., 1996). As demonstrated by Tanksley et al. (1989) with computer simulations, use of molecular markers for background selection can accelerate recovery of the RPG by two or three generations.

Background selection has two goals: (i) reduction of the proportion of the donor genome on the carrier chromosome of the target allele and (ii) reduction of the donor genome on the non-carrier chromosomes. The length of the chromosome segment from the donor that is linked to the target allele (linkage drag) is reduced by selecting individuals that carry the target allele and are homozygous for the RP alleles at tightly linked marker loci. In practical implementations of MAB, two crucial questions are How should the flanking markers by positioned? and How many individuals must be generated and genotyped with molecular markers to reduce the undesirable donor genome below a certain threshold?

Hospital et al. (1992) determined optimum distances d_1 and d_2 between the target locus and the flanking marker loci to recover a maximum amount of the RPG on the carrier chromosome by applying equation

Institute of Plant Breeding, Seed Science, and Population Genetics, Univ. of Hohenheim, 70593 Stuttgart, Germany. Received 31 March 1998. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 39:967–975 (1999).

Abbreviations: BC, backcross; BCr, r-th backcross generation; cM, centimorgan; MAB, marker-assisted backcrossing; NRP, non-recurrent parent; QTL, quantitative trait loci; RP, recurrent parent; RPG, recurrent parent genome; RFLP, restriction fragment length polymorphism.

LINEAR, BILINEAR, AND LINEAR-BILINEAR MODELS FOR ANALYZING GENOTYPE x ENVIRONMENT INTERACTION

J. Crossa[1], F.A. van Eeuwijk[2], P.L. Cornelius [3], and M. Vargas[4]

[1] **Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT) Apdo. Postal 6-641, 06600 Mexico DF, Mexico.[2] Wageningen University, Dep. of Plant Science, Laboratory of Plant Breeding, P.O.Box 386, 6700 AJ Wageningen, The Netherlands.[3] Dep. of Statistics and Dep. of Agronomy, University of Kentucky, Lexington, KY 40546-0091.[4] Universidad Autonoma de Chapingo, Chapingo, Mexico.**

Key Words: Least Squares, Singular Value Decompostition, Environmental and Genotypic Covariables.

Introduction

The presence of genotype x environment interaction (G×E) in agriculture is expressed either as inconsistent responses of some genotypes relative to others due to genotypic rank change or as changes in the absolute differences between genotypes without rank change. For the description of the mean response of genotypes over environments and for studying and interpreting G×E in agricultural experiments, three classes of models are commonly used: (1) linear models; (2) bilinear models, and (3) linear-bilinear models. One class of linear models, namely factorial regression (FR) models, and one class of bilinear models, namely partial least square (PLS) regression, allow incorporation of external environmental and genotypic covariables directly into the model.

Early approaches for analyses of G×E included the conventional fixed effects two-way (FE2W) model with sum to zero constraints running over indices. The empirical mean response, $\bar{y}_{ij}$, of the i^{th} genotype (i=1,2,...,I) in the j^{th} environment (j=1,2,...,J) with n replications in each of the I×J cells is expressed as

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + \bar{\varepsilon}_{ij} \qquad \text{Eq. 1}$$

where μ is the grand mean over all genotypes and environments, τ_i is the additive effect of the i^{th} genotype, δ_j is the additive effect of the j^{th} environment, $(\tau\delta)_{ij}$ is the non-additivity interaction (G×E) of the i^{th} genotype in the j^{th} environment and $\bar{\varepsilon}_{ij}$ is the average error assumed to be NID $(0, \sigma^2/n)$ (where σ^2 is the within-environment error variance, assumed to be constant). Yates and Cochran (1938) introduced the model in which the G×E term is linearly related to the environmental main effect.

The purpose of this paper is to present parsimonious approaches other than the FE2W model to the analysis of G×E. Examples illustrating the use of various statistical models for analyzing G×E in the context of plant breeding, genetics, and agronomy are given.

Linear-bilinear models

Williams (1952) was the first author to link the FE2W model with principal components (PC) analysis by considering the model

$\bar{y}_{ij} = \mu + \tau_i + \lambda\alpha_i\gamma_j + \bar{\varepsilon}_{ij}$ where λ is the largest singular value of $\mathbf{ZZ'}$ and $\mathbf{Z'Z}$ (for $\mathbf{Z} = \bar{y}_{ij} - \bar{y}_{i.}$) and α_i and γ_j are the corresponding eigenvectors.

Gollob (1968) and Mandel (1969, 1971) extended Williams' (1952) work by considering the bilinear G×E term as $(\tau\delta)_{ij} = \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk}$. Thus, the general formulation of the linear-bilinear model is

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij} \qquad \text{Eq. 2}$$

where the constant λ_k is the singular value of the k^{th} multiplicative component that is ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$; the α_{ik} are elements of the k^{th} left singular vector of the true interaction and represents the genotypic sensitivity to hypothetical environmental factors represented by the k^{th} right singular vector with elements γ_{jk}. The α_{ik} and γ_{jk} satisfy the ortho-normalization constraints $\sum_i \alpha_{ik}\alpha_{ik'} = \sum_j \gamma_{jk}\gamma_{jk'} = 0$ for $k \neq k'$ and $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$. When Eq. 2 is saturated the number of bilinear terms is t=min(I-1,J-1). Gabriel (1978) described the least squares fit of Eq. 2 and explained how the residual matrix of the G×E term, $\mathbf{Z} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$, is subjected to a singular value decomposition (SVD) after adjusting for the

additive (linear) terms. Gauch (1988) called the Eq. 2 Additive Main Effects and Multiplicative Interaction (AMMI) model.

Other classes of linear-bilinear models, described by Cornelius et al. (1996), are Genotypes Regression Model (GREG) $\bar{y}_{ij} = \mu_i + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$, the Sites (environments) Regression Model (SREG) $\bar{y}_{ij} = \mu_j + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$, the Completely Multiplicative Model (COMM) $\bar{y}_{ij} = \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$, and the Shifted Multiplicative Model (SHMM) $\bar{y}_{ij} = \beta + \sum_{k=1}^{t} \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$.

The SHMM model was the first linear-bilinear model used for identifying subsets of genotypes or environments in which genotypic rank changes would be negligible (Cornelius et al., 1992, 1993; Crossa et al., 1993, 1995, 1996; Crossa and Cornelius, 1993). The SREG model is useful in plant breeding because the bilinear terms contain both the main effects of genotypes and G×E. (Crossa and Cornelius, 1997).

In matrix notation, these linear-bilinear models can be expressed as $\mathbf{Y}=\sum_{k=1}^{m}\beta_k \mathbf{X}_k + \mathbf{AAG'}+\mathbf{E}$ (Cornelius and Seyedsadr, 1997) where $\mathbf{Y}=[\bar{y}_{ij}]$, $\mathbf{X}_k=[x_{kij}]$, $\mathbf{E}=[\bar{\varepsilon}_{ij}]$, $\mathbf{A}=\text{diag}(\lambda_k,\ k=1,2,\ldots,t)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$, $\mathbf{A}=(\alpha_1,\ldots,\alpha_t)$, $\mathbf{G}=(\gamma_1,\ldots,\gamma_t)$, and $\mathbf{A'A}=\mathbf{G'G}=\mathbf{I}_t$. The x_{kij} are known constants and β_k, λ_k, α_{ik}, and γ_{jk} are parameters to be estimated.

Linear models

The G×E is modeled directly using regression on environmental (and/or genotypic) variables. A useful linear model for incorporating external environmental (or genotypic) variables is the factorial regression (FR) model (Denis, 1988; van Eeuwijk et al., 1996). The FR models are ordinary linear models that approximate the G×E effects of Eq. 1 by the products of one or more of (1) genotypic covariables (observed) × environmental potentialities (estimated), (2) genotypic sensitivities (estimated) × environmental covariables (observed). For k=1,...,G genotypic covariables (centered) represented by $x_{i1},\ldots,x_{iG}$, Eq. 1 becomes $\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{g=1}^{G} x_{ig}\xi_{jg} + \bar{\varepsilon}_{ij}$, $G \leq J-1$, where ξ_{jg} represents an environmental factor (regression coefficient) with respect to the genotypic covariable, x_{ig}. Constraints on the parameters are $\sum_i \tau_i = \sum_j \delta_j = \sum_j \xi_{jg} = 0$. In matrix notation the expectation is

$$E(\mathbf{Y})=\mu\mathbf{1}_I\mathbf{1}'_J + \tau\mathbf{1}'_J + \mathbf{1}_I\delta' + \mathbf{X}\Xi' \qquad \text{Eq. 3}$$

where $\mathbf{Y}=[\bar{y}_{ij}]$ is a $I \times J$ matrix; $\mathbf{1}_I$ and $\mathbf{1}_J$ are $I \times 1$ and $J \times 1$ vectors with all elements equal to one, respectively; $\tau=[\tau_i]$ is the $I \times 1$ vector of main effects of genotypes; $\delta=[\delta_j]$ is the $J \times 1$ vector of main effects of environments; $\mathbf{X}=[x_{ig}]$ is the $I \times G$ matrix of known genotypic covariables; $\Xi=[\xi_{jg}]$ is the $J \times G$ matrix of unknown environmental factors (regression coefficients).

For h=1,..., H environmental covariables (centered) represented by $z_{j1},\ldots,z_{jH}$, Eq. 1 is $\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{h=1}^{H} \varsigma_{ih} z_{jh} + \bar{\varepsilon}_{ij}$, $H \leq J-1$, where ς_{ih} represents an genotypic sensitivity (regression coefficient) with respect to the to the environmental covariable, z_{jh}. Constraints on the parameters are $\sum_i \tau_i = \sum_j \delta_j = \sum_i \varsigma_{ih} = 0$. In matrix notation, the expectation is

$$E(\mathbf{Y})=\mu\mathbf{1}_I\mathbf{1}'_J + \tau\mathbf{1}'_J + \mathbf{1}_I\delta' + \zeta\mathbf{Z}' \qquad \text{Eq. 4}$$

where $\mathbf{Z}=[z_{jh}]$ is the $J \times H$ matrix of known environmental covariables; $\zeta=[\varsigma_{ih}]$ is the $I \times H$ matrix of unknown differential genotypic sensitivities.

The FR model including genotypic and environmental covariables simultaneously is $\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{g=1}^{G} x_{ig}\xi_{jg} + \sum_{h=1}^{H} \varsigma_{ih} z_{jh} + \sum_{h=1}^{H} \sum_{g=1}^{G} x_{ig} v_{gh} z_{jh} + \bar{\varepsilon}_{ij}$ where v_{kh} is a constant that scales the cross-product of the genotypic covariables, x_k, with the environmental covariables, z_h, and can be derived from the two previous FR models by imposing the restriction $\xi_{jg} = v_{gh} z_{jh}$ or $\varsigma_{ih} = x_{ih} v_{gh}$; each

cross product represents one degree of freedom in the G×E subspace. In matrix notation the expectation is $E(\mathbf{Y})=\mu\mathbf{1}_I\mathbf{1}'_J+\tau\mathbf{1}'_I+\mathbf{1}_J\delta'+\mathbf{X}\nu\mathbf{Z}+\mathbf{X}\Xi'+\zeta\mathbf{Z}'$ where the constraint $\mathbf{X}\Xi'=\zeta\mathbf{Z}'=\mathbf{0}$ (where $\mathbf{0}$ is a matrix H×G of zeros) is required. The model should be fitted for all possible combinations of genotypic covariables with environmental covariables.

When environmental (or genotypic) covariables show high collinearity, interpretation of the least squares regression coefficients is complicated because they are estimated very imprecisely. Consequently, a stepwise procedure for choice of the covariables to include could be useful for model construction. Noise on the response variable also complicates the interpretation of the FR parameters. Furthermore, least squares estimation of the parameters in the FR models are not unique when the number of covariables is larger than the number of observations, so an alternative estimation method is needed. Partial Least Squares (PLS) regression overcomes some of these problems and it can be used as an alternative estimation method.

Bilinear models

Multivariate Partial Least Squares (PLS) regression models (Aastveit and Martens, 1986; Helland, 1988) are a special class of bilinear models. When genotypic responses over environments ($\mathbf{Y}$) are modeled using environmental covariables, then the $J\times H$ matrix $\mathbf{Z}$ of H (h=1,2,...,H) environmental covariables can be written in a bilinear form as

$$\mathbf{Z}=\mathbf{t}_1\mathbf{p}'_1+\mathbf{t}_2\mathbf{p}'_2+\ldots+\mathbf{t}_M\mathbf{p}'_M+\mathbf{E}_M=\mathbf{TP}'+\mathbf{E} \qquad \text{Eq. 5}$$

where the matrix T contains the $\mathbf{t}_1$ $J\times 1$ vectors called latent environmental covariables or Z-scores (indexed by environments) and the matrix $\mathbf{P}$ has the $\mathbf{p}_1\ldots\mathbf{p}_H$ $H\times 1$ vectors called Z-loadings (indexed by environmental variables) and $\mathbf{E}$ has the residuals. Similarly, the response variable matrix $\mathbf{Y}$ in bilinear form is

$$\mathbf{Y}=\mathbf{t}_1\mathbf{q}'_1+\mathbf{t}_2\mathbf{q}'_2+\ldots+\mathbf{t}_M\mathbf{q}'_M+\mathbf{F}_M=\mathbf{TQ}'+\mathbf{F} \qquad \text{Eq. 6}$$

where the matrix T is as in Eq. 5 and the matrix $\mathbf{Q}$ contains the $\mathbf{q}_1\ldots\mathbf{q}_I$ $I\times 1$ vectors called Y-loadings (indexed by genotypes) and $\mathbf{F}$ has the residuals. The relationship between $\mathbf{Y}$ and $\mathbf{Z}$ is transmitted through the latent variable T. The PLS algorithm performs separate (but simultaneous) principal component analysis of $\mathbf{Z}$ and of $\mathbf{Y}$ that allows reduction of the number of variables in each system to a smaller number of hopefully more interpretable latent variables.

Helland (1988) showed that a reduced number of PLS latent variables gives a low rank representation of the least squares estimates of the FR with environmental covariables because the expectation of $\mathbf{Y}'$ is

$$E(\mathbf{Y}')=\mathbf{QT}'=\mathbf{Q}(\mathbf{ZW})'=(\mathbf{QW}')\mathbf{Z}'=\zeta\mathbf{Z}'=\sum_{h=1}^{H}\zeta_{ih}z_{jh} \qquad \text{Eq. 7}$$

as in Eq. 4 where T, $\mathbf{Q}$, and $\mathbf{Z}$ are defined as before and the vector $\mathbf{W}$ is $H\times 1$ and contains the Z-loadings (or weights) of the environmental covariables; ζ contains the PLS approximation to the regression coefficients of the responses in $\mathbf{Y}$ to the environmental covariables in $\mathbf{Z}$. The matrices T (with J coordinates for environments), $\mathbf{Q}$ (with I coordinates for genotypes) and $\mathbf{W}$ (with H coordinates for environmental covariables) can be represented in the PLS biplot such that projecting the j^{th} environment (row) of T on the i^{th} genotype (row) of $\mathbf{Q}$ [$\mathbf{Y}'=(\mathbf{TQ}')'$] approximates the G×E; projecting the h^{th} environmental covariable (row) of $\mathbf{W}$ on the i^{th} genotype (row) of $\mathbf{Q}$ ($\mathbf{QW}'=\zeta$) approximates the regression coefficient of the i^{th} genotype on the h^{th} environmental covariable (Vargas et al.,1999; van Eeuwijk et al., 2000). When genotypic covariables are used to model environmental responses over genotypes, then the latent genotypic covariables are $\mathbf{T}=\mathbf{XW}$ where vector $\mathbf{W}$ is $G\times 1$ and contains the weights of the genotypic covariables. The expectation of $\mathbf{Y}$ is

$$E(\mathbf{Y})=\mathbf{TQ}'=\mathbf{XWQ}'=\mathbf{X}\Xi'=\sum_{g=1}^{G}x_{ig}\xi_{jg} \qquad \text{Eq. 8}$$

as in Eq. 3 (van Eeuwijk et al., 2000; Vargas et al., 1999) where Ξ contains the PLS approximation to the regression coefficients of the responses in $\mathbf{Y}$ to the genotypic covariables in $\mathbf{X}$. The matrices T (with I coordinates for genotypes), $\mathbf{Q}$ (with J coordinates for environments) and $\mathbf{W}$ (with G coordinates for genotypic covariables) can be represented in a PLS biplot such that projection of the i^{th} genotype (row) of T onto the j^{th} environment (row) of $\mathbf{Q}$ ($\mathbf{Y}=\mathbf{TQ}'$) approximates the G×E; projection the g^{th} genotypic covariable (row) of $\mathbf{W}$ onto the j^{th} environment (row) of $\mathbf{Q}$ ($\mathbf{WQ}'=\Xi$) approximates the regression coefficient of the j^{th} environment on the g^{th} genotypic covariable.

QTL and QTL× environment interaction analysis in genetics and plant breeding

In plant breeding much research is directed at locating the regions of the chromosomes that are involved in the physiological processes underlying phenotypical traits. These regions are called quantitative trait loci (QTL or QTLs). When these regions differ between genotypes in relation to changes in the environment, QTL×environment interaction occurs (QTL×E). The statistical problem can be interpreted as a multivariate multiple regression of phenotypic traits as observed over a set of environments on a set of genetic predictors. FR provides a suitable framework for QTL×E analysis. In

Crossa et al. (1999) examples are given of how FR and PLS can be used for assessing location and importance of QTL and QTL×E.

FR models of the form

$\bar{y}_{ij} = \mu + \delta_j + \sum_{g=1}^{G} x_{ig}\xi_g + \bar{\varepsilon}_{ij}$ and

$\bar{y}_{ij} = \mu + \delta_j + \sum_{g=1}^{G} x_{ig}\xi_{jg} + \bar{\varepsilon}_{ij}$ (van Eeuwijk, et al., 2000) are useful for studying QTL and QTL×E, respectively where x_{ig}'s are genotypic covariables, or genetic predictors, at specific locations of the chromosmes, whose values are functions of the neighboring genetic markers and the position at the chromosome. The ξ_g's represent the QTL main effects, which are indexed by environment, ξ_{jg}, represents the QTL×E. Following Haley and Knott (1992), the simplest QTL mapping analysis considers the regression on the genetic predictors at marker positiuns (individual marker regression) where the additive effects of the observed marker genotypes *MM*, *Mm* and *mm* are 1,0, and –1, respectively and the dominance effects for *MM*, *Mm* and *mm* are 0,1, and 0, respectively. Somewhat more advanced, simple interval QTL mapping analysis cunsiders the regression on the genetic predictors not only at marker positions but also at regular intervals between markers. The additive effect and dominance effects can be computed.

In composite interval QTL mapping analysis a correction is added for the effects of QTLs at other positions in the genome. Let the position under evaluation be *p*, then other markers, called cofactors, *C*, are included in the model to reduce noise created by the effect of other QTLs, then the model is $\bar{y}_{ij} = \mu + \delta_j + \sum_{g \in C} x_{ig}\xi_g + x_{ip}\xi_p + \bar{\varepsilon}_{ij}$. Selection of the appropriate markers to be used as cofactors for correcting the effect of other QTL can be done by one of a few PLS-axes created by regressing the multivariate response on all genetic predictors outside the evaluation window in composite interval mapping, and then perform the mapping procedure with the corrected responses. Testing procedures for the presence of QTL and QTLxE at a certain position can be done by permutation tests (van Eeuwijk, at al., 2000).

Some results of the application of the methodology described in van Eeuwijk et al. (2000) now follow. The grain yield of F_2 (211) tropical CIMMYT maize lines lines was evaluated in eight environments that were contrasting in drought and nitrogen stress. As FR is essentially a regression method, QTLs and QTL×E can be located by the application of standard F-statistics. Plot of F-profiles over the first chromosome is in Fig. 1. Based on randomization studies thresholds should be applied of about 54, 4.5 and 9, respectively (α=0.05).

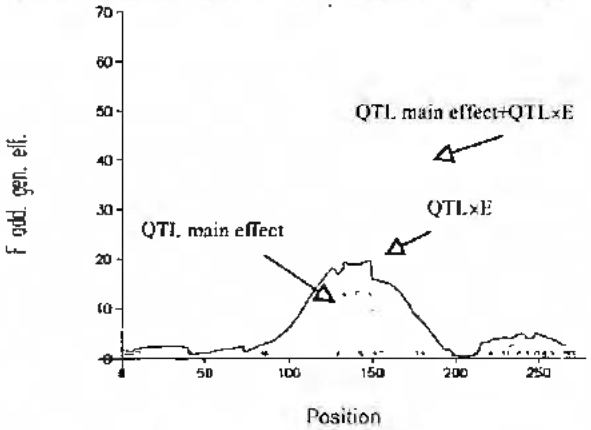

Figure 1. F-values for the different positions along the first chromosome for additive genetic QTL main effects, additive genetic QTL×E, and QTL main effects+QTL×E.

Treatment × environment interaction analysis in agronomy

A parsimonious description of the Treatment × Environment (T×E) existing in 24 agronomic treatments (1-24) [tillage, summer crop, manure, and nitrogen (N)] evaluated during 10 consecutive years (1988-97) was conducted by Vargas et al. (2001). Results of the final MFR were compared with those of a Partial Least Squares (PLS) to achieve extra insight in both the T×E and the final multiple factorial regression (MFR) model.

The MFR was applied on the six most important components of the T×E term: Year×Tillage, Year×Summer Crop, Year×Manure, Year×N, Year×Summer Crop×N, and Year× Manure× N. Results for the MFR of the 27 environmental covariables × tillage interaction showed that evaporation in December (EVD) × tillage sum of squares accounted for 68% of the whole year × tillage interaction. For year × summer crop, evaporation in April (EVA) accounted for 36% of the year × summer crop. For year × manure, covariables precipitation in December (PRD) and sun hours in February (SHF) contributed with 56% of the year × manure sum of squares. Year × Nitrogen (N) interaction determined the major part of year × treatment interaction sum of squares.

The PLS biplot separated the nine highest yielding treatments (9,19,21,17,11,12,10,23, and 18) from the nine lowest yielding treatments (1,2,3,4,5,6,7,8, and 16) (Fig. 3). The nine lowest yielding treatments had a

positive interaction with year 1995 that had high mTUF, mTF, and MTA but a negative interaction with year 1988 (opposite quadrant). The PLS biplot contains roughly five clusters of correlated environmental covariables. The order of inclusion of these covariables in the MFR with the stepwise procedure for each factor effect corresponds to selecting covariables for the different cluster groups depicted in Fig. 3.

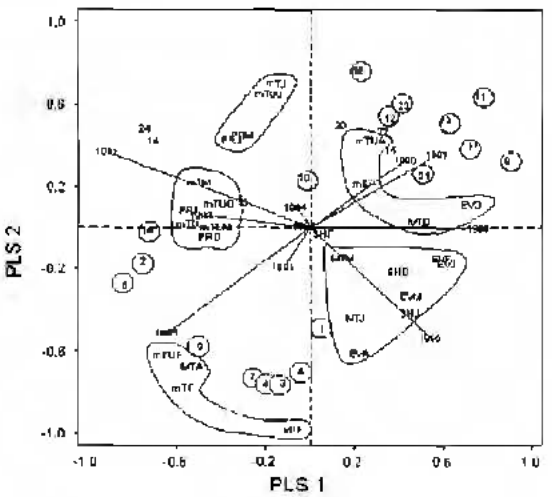

Figure 3. Biplot of the first and second PLS factors representing the Z-scores of the 10 years (1988-97), and the Y-loadings of the 24 practice treatments (1-24) enriched with the Z-loadings of 27 environmental variables: EV: total monthly evaporation, PR: total monthly precipitation, SH: sun hours per day, mT: mean minimum temperature sheltered, MT: mean maximum temperature sheltered, mTU: mean minimum temperature unsheltered; D: December, J: January, F: February, M: March, A: April; N: Nitrogen (from Vargas et al., 2001).

Crossover interaction analysis in plant breeding

Using linear-bilinear SHMM model, Cornelius et al. (1992) defined sufficient conditions for the absence of significant genotype crossover interaction (COI) in a set of environments and/or genotypes: (1) SHMM with t=1 ($SHMM_1$) must be an adequate model for fitting the data and (2) $\hat{\gamma}_{j1}$ are all of like sign. When $SHMM_1$ predicted values, $\hat{y}_{ij} = \hat{\beta} + \hat{\lambda}_1 \hat{\alpha}_{i1} \hat{\gamma}_{j1}$, are plotted against the primary effects of environments, $\hat{\gamma}_{j1}$, the graph consists of a set of regression lines, one for each genotype, all of which concur at the point (0, $\hat{\beta}$). For a non-COI $SHMM_1$, the $\hat{\gamma}_{j1}$ are all of like sign (or zero) and, thus, the point of intersection is a point either at the boundary (if one $\hat{\gamma}_{j1}$=0), or outside (left or right) of the region containing the plotted points. If the $\hat{\gamma}_{j1}$ have different signs, then the point of concurrence is within the region containing the plotted points and a complete reversal of rank order of genotypes is displayed on the right, as compared to the left, of the point of concurrence. For clustering environments, $SREG_1$ can be used instead of $SHMM_1$ and all the above properties still hold then the figure shows an overlayed set of broken lines (one for each genotype) that display no genotype COI within the region of plotted points.

When SHMM is fitted to the entire set of data, several components are necessary if an adequate fit is to be achieved. The procedure by a which subset of environments without COI is found consists in using a clustering strategy that will divide the environments into subsets such that significant variation captured as secondary, tertiary, etc., effects when SHMM is fitted to the entire data set, can be expressed as primary effects in separate analyses of data from the subsets. The measure of distance between two environments is the residual mean square after fitting $SHMM_1$ [$RMS(SHMM_1)$] to the data from the two environments subject to a non-COI constraint.

Data from a trial with g=9 genotypes evaluated in e=20 environments showed that $SHMM_1$ will not adequately fit the entire data and the fitted $SHMM_1$ itself displayed genotype COI. In Fig. 4, the consistant response of the nine genotypes across a subset of ten environments is depicted through the overlayed broken line $SREG_1$ that does not cross over.

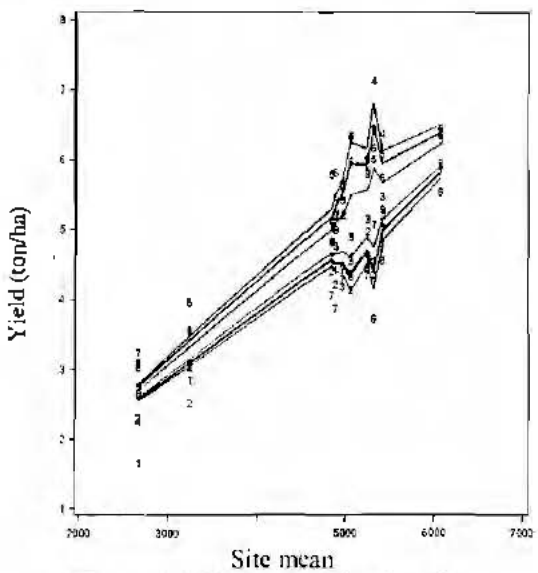

Figure 4. $SREG_1$ model fitted to nine genotypes and a subset of ten environments.

References

Aastveit, H. and H. Martens (1986). Anova interactions interpreted by partial leats squares regression. Biometrics 42:829-844.

Cornelius, P.L, Seyedsadr, M., and Crossa, J. (1992). Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theoretical and Applied Genetics 84:161-172.

Cornelius, P.L.,Van Sanford, D.A, and Seyedsadr, M. (1993). Clustering cultivars into groups without rank-change interactions. Crop Science 33:1193-1200.

Cornelius, P.L., Crossa, J., and Seyedsadr, M.(1996). Statistical tests and estimators for multiplicative models for cultivar trials. In Kang, M.S., and Gauch, H.G., Jr., (Eds.), Genotype-by-Environment Interaction. Boca Raton: CRC Press, pp. 199-234.

Cornelius, P.L., and Seyedsadr, M. (1997). Estimation of general linear-bilinear models for two-way tables. Journal of Statistical Computation and Simulation 58: 287-322.

Crossa, J., Cornelius P.L., Seyedsadr, M., and Byrne, P. (1993). A shifted multiplicative model cluster analysis for grouping environments without genotypic rank-change. Theoretical and Applied Genetics 85:577-586.

Crossa, J., and Cornelius, P.L. (1993) Recent developments in multiplicative models for cultivar trials. In: Buxton, D.R., et al. (Eds.) International Crop Science I. Crop Science Society of America Madison, Wisconsin.

Crossa, J., Cornelius, P.L., Sayre, K., and Ortiz-Monasterio I.J. (1995). A shifted multiplicative model fusion method for grouping environments without cultivar rank change. Crop Science 35:54-62.

Crossa, J., Cornelius, P.L., and Seyedsadr, M.(1996). Using the shifted multiplicative model cluster methods for crossover genotype-by-environment interaction. In Kang, M.S. and Gauch, H.G., Jr., (Eds.), Genotype-by-Environment Interaction. Boca Raton: CRC Press, pp. 175-198.

Crossa, J., and Cornelius, P.L. (1997). Sites regression and shifted multiplicative model clustering of cultivar trials sites under heterogeneity of error variance. Crop Science 37: 405-415.

Crossa, J., M. Vargas, F.A. van Eeuwijk, C. Jiang, G.O. Edmeades, and D. Hoisington. (1999). Interpreting genotype×environment interaction in tropical maize using linked molecular markers and environmental covariables. Theoretical and Applied Genetics 99:611-625.

Denis, J-B. (1988). Two-way analysis using covariates. Statistics 19:123-132.

Gabriel, K.R. (1978). Least squares approximation of matrices by additive and multiplicative models. Journal of the Royal Statistical Society, Series B, 40:186-196.

Gauch, H.G. Jr. (1988). Model selection and validation for yield trials with interaction. Biometrics 44:705-715.

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

Helland, I.S. (1988). On the structure of partial least squares. Communications in Statistics, Part B Simulations and Computations 17:581-607.

Haley, C.S. and S.A. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324.

Mandel, J. (1969). The partitioning of interaction in analysis of variance. Journal of Research of the National Bureau of Standards, Series B, 73:309-328.

Mandel, J. (1971). A new analysis of variance model for non-additive data. Technometrics 13:1-18.

Vargas, M., J. Crossa, F.A. van Eeuwijk, M.E. Ramirez and K. Sayre. (1999). Using partial Least Squares, Factorial Regression and AMMI models for interpreting Genotype×Environment Interaction. Crop Science 39:955-967.

Vargas, M., J. Crossa, F.A. van Eeuwijk, K. Sayre, and M.P. Reynolds. (2001). Interpreting Treatment×Environment interaction in Agronomy Trials. Agronomy Journal (in press).

van Eeuwijk, F.A., J-B. Denis, and M.S. Kang. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In S. Kang and H.G. Gauch (eds). Genotype-by-environment interaction, CRC Press, Boca Raton, Fl.

van Eeuwijk, F.A., J. Crossa, M. Vargas, J.M. Ribaut (2000). Variants of factorial regresión for analysing QTL by environment interaction. Proceedings of the 11th Meeting of the EUCARPIA Section Biometrics in Plant Breeding. In A. Gallais, C. Dillmann, and I. Goldringer (eds). Quantitative genetics and breeding methods: the way ahead. Paris, France.

Williams, E.J. (1952) The interpretation of interactions in factorial experiments. Biometrika 39:65-81.

Yates, F., and Cochran, W.G. (1938). The analysis of groups of experiments. Journal of Agricultural Science 28:556-580.

ANEXO 2
MATERIAL ENTREGADO A LOS PARTICIPANTES DEL CURSO

GOBIERNO DE CHILE
FUNDACIÓN PARA LA
INNOVACIÓN AGRARIA

ANEXO 3
CARTA DE AGRADECIMIENTO AL DIRECTOR DE CIMMYT

Santiago, November 21, 2004.

Dr. Masaru Iwanaga
Director General
CIMMYT
Apdo Postal 6-641
06600 Mexico DF
Mexico

Dear Dr. Iwanaga,

This is to acknowledge the recent participation of Dr. José Crossa as an Invited Professor to the University of Chile. Dr. Crossa spent a week with us in Santiago and offered a Course in Selected Topics of Experimental Design and Data Analysis with Emphasis on Multivariate Analysis. Twenty five students including University Professors, Graduate Students, INIA Researchers and Research Staff of major Seed Production Companies attended the course.

This activity turned to be of major importance to upgrade our knowledge in these matters, particularly considering that Dr. Crossa is one of the world leading scientists in the subject. His course was rated excellent by all participants. The usefulness of the topics treated for agriculture as well as for other research areas is beyond expectations. A highly complex subject was taught in an elegant conceptual way, such that all participants understood the principles involved. At the same time, continuous, carefully chosen readings and exercises put the concepts to work. A final comprehensive take home exam, approved by all participants, was the final test for this successful project. The Course was submitted to the Graduate School of the Agronomy Faculty of our University for evaluation, such that it would provide credit to the graduate students who attended and approved it.

Financial support for Dr. Crossa's air travel and expenditures while in Chile, as well as for minor expenditures related to the course, was provided by the Chilean National Fund for Innovation in Agriculture (FIA) that belongs to the Chilean Ministry of Agriculture. The Course preparation was carried out by the Soil-Plant-Water Relations Laboratory of the Faculty of Agronomy of the University of Chile. The computer hardware as well as other infrastructure was provided by the Faculty of Agronomy. Dr. Crossa brought with him the specially required computer software.

Finally, I would like to express my gratitude to Dr. José Crossa for graciously sharing with us his valuable knowledge and experience as well as.my recognition to CIMMYT for maintaining state of the art research in this essential area for agricultural development.

Sincerely yours,

Edmundo Acevedo
Ing. Agr. MS PhD
Profesor Titular
Universidad de Chile

cc.
Decano Facultad de Agronomía
Deputy Director of Research CIMMYT
Director de Programa de Doctorado CSAV