



CIMMYT®

Linear, bilinear, and linear-bilinear models for analyzing $G \times E$ in plant breeding and agronomy research

Jose Crossa

Biometrics and Statistics Unit

International Maize and Wheat Improvement Center (CIMMYT)

Apdo. Postal 6-641, 06600 Mexico DF, Mexico.

J.CROSSA@CGIAR.ORG



Basic ANOVA model

- ◆ The response of the i^{th} genotype in the j^{th} environment is

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + (\tau\delta)_{ij} + \bar{\varepsilon}_{ij}$$

This model is

- ▶ Unparsimonious each $G \times E$ cell has its own interaction parameter
- ▶ Uninformative the independent interaction parameters are difficult to interpret.

History of linear-bilinear models for assessing GxE

- ◆ Yates and Cochran (1938) $(\tau \delta)_{ij} = \xi_i \delta_j$
- ◆ Tukey (1949) proposed testing $K=0$ in the model $(\tau \delta)_{ij} = K \tau_i \delta_j$
- ◆ Williams (1952) considers $\bar{y}_{ij} = \mu + \tau_i + \lambda \alpha_i \gamma_j + \bar{\varepsilon}_{ij}$ where λ is the largest singular value of \mathbf{ZZ}' and $\mathbf{Z}'\mathbf{Z}$ ($\mathbf{Z} = \bar{y}_{ij} - \bar{y}_i$) and α_i and γ_j are the eigenvectors
- ◆ Mandel (1961) generalized Tukey's model
 - ▶ $(\tau \delta)_{ij} = \lambda \alpha_i \delta_j$ for genotypes or
 - ▶ $(\tau \delta)_{ij} = \lambda \tau_i \gamma_j$ for environments.

- ◆ Gollob (1968) and Mandel (1969, 1971) extended Williams' (1952) work such that $(\tau \delta)_{ij} = \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk}$

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

Gabriel (1978) described the least squares fit of the G×E term $z_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$

Zobel et al. (1988) and Gauch (1988) called AMMI

Other classes of linear-bilinear models (Cornelius et al., 1996)

Genotypes Regression Model (**GREG**)

$$\bar{y}_{ij} = \mu_i + \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

Sites Regression Model (**SREG**)

$$\bar{y}_{ij} = \mu_j + \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

Completely Multiplicative Model (**COMM**)

$$\bar{y}_{ij} = \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

Shifted Multiplicative Model (**SHMM**)

$$\bar{y}_{ij} = \beta + \sum_{k=1}^t \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

- ◆ The SHMM model is used for identifying subsets of genotypes or environments with negligible COI.
- ◆ The SREG model has been used in preference to SHMM for grouping environments without genotypic rank change.

Linear Model

Factorial regression (FR) Model

Objective

To replace, in the $G \times E$ subspace, the genotypic and environmental factors by a small number of genotypic and environmental covariables (Denis, 1988; van Eeuwijk et al., 1996).

- ◆ The FR models are ordinary linear models that approximate the $G \times E$ effects by the products of one or more of
 - (1) genotypic covariables (observed) \times environmental potentialities (estimated),
 - (2) genotypic sensitivities (estimated) \times environmental covariables (observed),

Factorial regression Model

ENVIRONMENTAL COVARIABLES

For $h=1, \dots, H$ environmental covariables (centered), z_{j1}, \dots, z_{jH} , ($H \leq J-1$)

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{h=1}^H \zeta_{ih} z_{jh} + \bar{\varepsilon}_{ij}$$

ζ_{ih} represents regression coefficient with respect to the environmental covariable z_{jh}

Sum to zero constraints on the parameters are

$$\sum_i \tau_i = \sum_j \delta_j = \sum_i \zeta_{ih} = 0$$

In matrix notation, the expectation is $E(Y) = \mu \mathbf{1}_I \mathbf{1}'_J + \boldsymbol{\tau} \mathbf{1}'_J + \mathbf{1}_I \boldsymbol{\delta}' + \boldsymbol{\zeta} \mathbf{Z}'$

Factorial regression Model

GENOTYPIC COVARIABLES

For $k=1, \dots, G$ genotypic covariables (centered), x_{i1}, \dots, x_{ig} , ($G \leq I-1$)

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{g=1}^G x_{ig} \xi_{jg} + \bar{\varepsilon}_{ij}$$

ξ_{jg} represents regression coefficient with respect to the genotypic covariable x_{ig}

Constraints on the parameters are $\sum_i \tau_i = \sum_j \delta_j = \sum_i \xi_{ih} = 0$

In matrix notation, the expectation is $E(Y) = \mu \mathbf{1}_I \mathbf{1}'_J + \boldsymbol{\tau} \mathbf{1}'_J + \mathbf{1}_I \boldsymbol{\delta}' + \mathbf{X} \boldsymbol{\Xi}'$

Factorial regression Model

GENOTYPIC AND ENVIRONMENTAL COVARIABLES

For $k=1, \dots, G$ genotypic covariables (centered), $x_{i1}, \dots, x_{ig}, (G \leq I-1)$

$$\bar{y}_{ij} = \mu + \tau_i + \delta_j + \sum_{g=1}^G x_{ig} \xi_{jg} + \sum_{h=1}^H \zeta_{ih} z_{jh} + \sum_{g=1}^G \sum_{h=1}^H x_{ig} v_{gh} z_{jh} + \bar{\varepsilon}_{ij}$$

where v_{gh} is a constant that scales the cross-product of the genotypic covariables x_k with the environmental covariables Z_h

Constraints on the parameters are $\xi_{jg} = v_{gh} z_{jh}$ or $\zeta_{ih} = x_{ih} v_{gh}$

In matrix notation, the expectation is $E(Y) = \mu \mathbf{1}_I \mathbf{1}'_J + \tau \mathbf{1}'_J + \mathbf{1}_I \delta' + \mathbf{X} \mathbf{v} \mathbf{Z} + \mathbf{X} \boldsymbol{\Xi}' + \boldsymbol{\zeta} \mathbf{Z}'$

where the constraint $\mathbf{X} \boldsymbol{\Xi}' + \boldsymbol{\zeta} \mathbf{Z}' = \mathbf{0}$ is required ($\mathbf{0}$ is a matrix $H \times G$ of zeros)

Problems with Factorial Regression

- When environmental (or genotypic) covariables show **high collinearity**, interpretation of the least squares regression coefficients is complicated because they are estimated very imprecisely.
- Noise on the response variable also complicates the interpretation of the FR parameters.
- 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 ($p \gg n$)

SOLUTION

Consequently, a stepwise procedure for choice of the covariables to include could be useful for model construction. 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 Model

Partial Least Squares (PLS)

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 ($h=1,2,\dots,H$) environmental covariables can be written in a bilinear form as

$$\mathbf{Z} = \mathbf{t}_1 \mathbf{p}_1' + \mathbf{t}_2 \mathbf{p}_2' + \dots + \mathbf{t}_M \mathbf{p}_M' + \mathbf{E}_M = \mathbf{TP}' + \mathbf{E}$$

where

- \mathbf{T} contains the \mathbf{t}_j $J \times 1$ vectors called **latent environmental covariables** or **Z-scores** (indexed by environments)
- \mathbf{P} has the $\mathbf{p}_1 \dots \mathbf{p}_H$ $H \times 1$ 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

$$\mathbf{Y} = \mathbf{t}_1 \mathbf{q}'_1 + \mathbf{t}_2 \mathbf{q}'_2 + \dots + \mathbf{t}_M \mathbf{q}'_M + \mathbf{F}_M = \mathbf{TQ}' + \mathbf{F}$$

where

→ The matrix \mathbf{Q} contains the $\mathbf{q}_1 \dots \mathbf{q}_I$ 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 \mathbf{T} .

The PLS algorithm performs separate (but simultaneous) principal component analysis of Z and of Y that allows reduction of the number of variables in each system to a smaller number of hopefully more interpretable latent variables.

PLS with environmental covariables

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 Y' is

$$E(Y') = QT' = Q(ZW)' = (QW')Z' = \zeta Z' = \sum_{h=1}^H \zeta_{ih} z_{jh}$$

- 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 Y to the environmental covariables in Z .

PLS Biplot with environmental covariables

Matrices

→ **T** (with J coordinates for environments),

→ **Q** (with I coordinates for genotypes) and

→ **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 **Q** [$\mathbf{Y}' = (\mathbf{TQ}')'$] approximates the $G \times E$;
- ◆ projecting the h th environmental covariable (row) of **W** on the i th genotype (row) of **Q** ($\mathbf{QW}' = \boldsymbol{\zeta}$) approximates the regression coefficient of the i th genotype on the h th environmental covariable

PLS with genotypic covariables

When genotypic covariables are used to model environmental responses over genotypes, then the latent genotypic covariables are $T=XW$ where vector W is $G \times 1$ and contains the weights of the genotypic covariables.

The expectation of Y is

$$E(Y)=TQ'=XWQ'=X\Xi'=\sum_{g=1}^G x_{ig} \xi_{jg}$$

→ Ξ contains the PLS approximation to the regression coefficients of the responses in Y to the genotypic covariables in X

PLS with genotypic covariables

Matrices

→ \mathbf{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 \mathbf{T} onto the j th environment (row) of \mathbf{Q} ($\mathbf{Y}=\mathbf{TQ}'$) approximates the $G \times E$;
- ◆ of the g th genotypic covariable (row) of \mathbf{W} onto the j th environment (row) of \mathbf{Q} ($\mathbf{WQ}'=\mathbf{\Xi}$) approximates the regression coefficient of the j th environment on the g th genotypic covariable.

**APPLICATIONS OF
FACTORIAL REGRESSION
AND PARTIAL LEAST
SQUARES IN AGRONOMY
RESEARCH**

Treatment \times environment interaction analysis in agronomy

24 agronomic treatments evaluated during 10 consecutive years.

Objectives

- (1) To investigate the factorial structure of the treatments to reduce the number of treatment terms in the interaction,
- (2) To use quantitative year covariables to replace the qualitative variable year and use multiple factorial regression (MFR) with a stepwise variable selection procedure for finding the most relevant environmental covariables.

Results of the final MFR were compared with those of PLS based analysis to achieve extra insight in both the T \times E and the final MFR model.

Table 2. Factorial regression model including only the six highly significant interaction terms and the variables found by stepwise for each factorial effect.

Source	df	Sum of Squares ($\times 10^6$)	Mean Squares ($\times 10^5$)	Prob
Treatment	23	773.970	336.508	0.0001
Year	9	373.260	414.733	0.0001
Year \times treatment	207	279.520	13.503	0.0001
Year \times tillage	9	21.070	23.410	0.0001
EVD\timestillage	1	14.290	142.900	0.0001
Deviations	8	6.780	8.480	0.0001
Year \times summer crop	9	8.729	9.699	0.0001
EVA\timessum crop	1	3.152	31.500	0.0001
Deviations	8	5.577	6.971	0.0001
Year \times manure	9	37.556	41.730	0.0001
PRD\timesmanure	1	16.170	161.700	0.0001
SHF\timesmanure	1	4.756	47.560	0.0001
Deviations	7	16.630	23.750	0.0001
Year \times N	18	126.900	70.500	0.0001
mTF\timesN	2	61.360	306.800	0.0001
mTJ\timesN	2	20.840	104.200	0.0001
MTA\timesN	2	25.580	127.900	0.0001
mTM\timesN	2	11.790	58.950	0.0001
Deviations	10	7.330	7.330	0.0009
Year \times sum crop \times N	18	18.325	10.180	0.0001
MTF\timessum crop\timesN	2	8.487	42.430	0.0001
Deviations	16	9.838	6.149	0.0008
Year \times manure \times N	18	31.366	17.430	0.0001
mTUM\timesmanure\timesN	2	19.050	95.250	0.0001
SHJ\timesmanure\timesN	2	5.457	27.290	0.0001
Deviations	14	6.859	4.899	0.0141
Error	460	110.870	2.410	

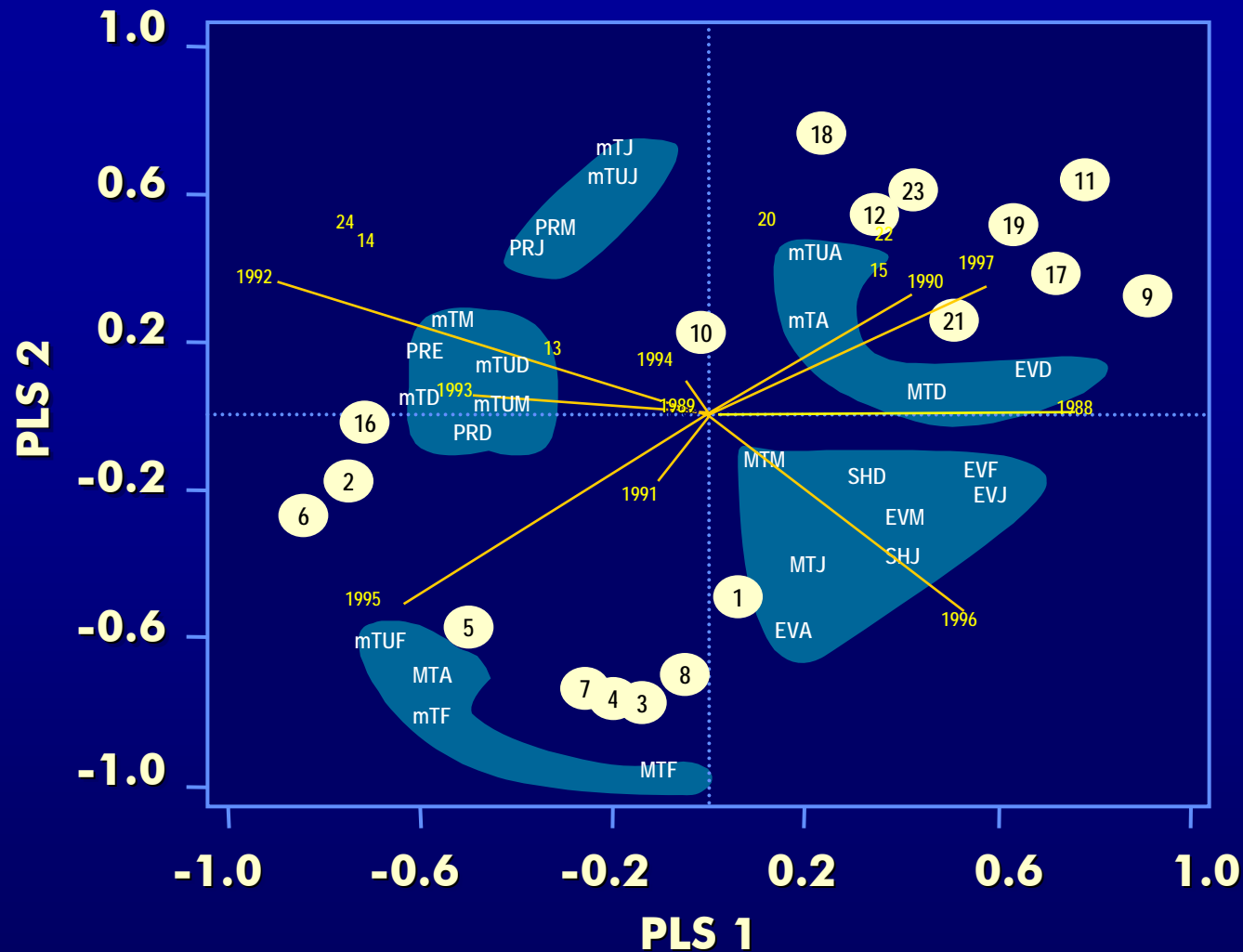


Figure 3. Biplot of the first and second PLS factors representing the **Z-scores** of the 10 years (1988-97), and the **Y-loadings** of the 24 practice treatments (1-24) enriched with the **Z-loadings** of 27 environmental variables: EV: total monthly evaporation, PR: total monthly precipitation, SH: 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).

Genotype x environment interaction for zinc and iron concentration of wheat grain in eastern Gangetic Plains of South Asia

20 elite CIMMYT wheat lines were evaluated in a multilocation trial in the Eastern Gangetic Plains (EGP) of India to determine GE for agronomic and nutrient traits.

Grain yield was available for 14 environments, while zinc and iron concentration of grains for 10 environments.

Soil and climatic data of each of the locations were also used

Environmental variables

TMXBF = Temperature maximum before flowering;

TMXAF = Temperature maximum after flowering;

TMNBF = Temperature minimum before flowering;

TMNAF = Temperature minimum after flowering;

RHBF = Relative humidity before flowering;

RHAF = Relative humidity after flowering;

RBF = Rainfall before flowering (RBF);

RAF = Rainfall after flowering (RAF).

Soil variables

Zn_30 = Zinc concentration in 0-30 cm soil depth;

Zn_60 = Zinc concentration in 30-60 cm soil depth;

Fe_30 = Iron concentration in 0-30 cm soil depth;

Fe_60 = Iron concentration in 30-60 cm soil depth.

Proportion of variation accounted from Factorial Regression Analysis for each of the significant covariables for grain yield, grain iron and zinc concentration

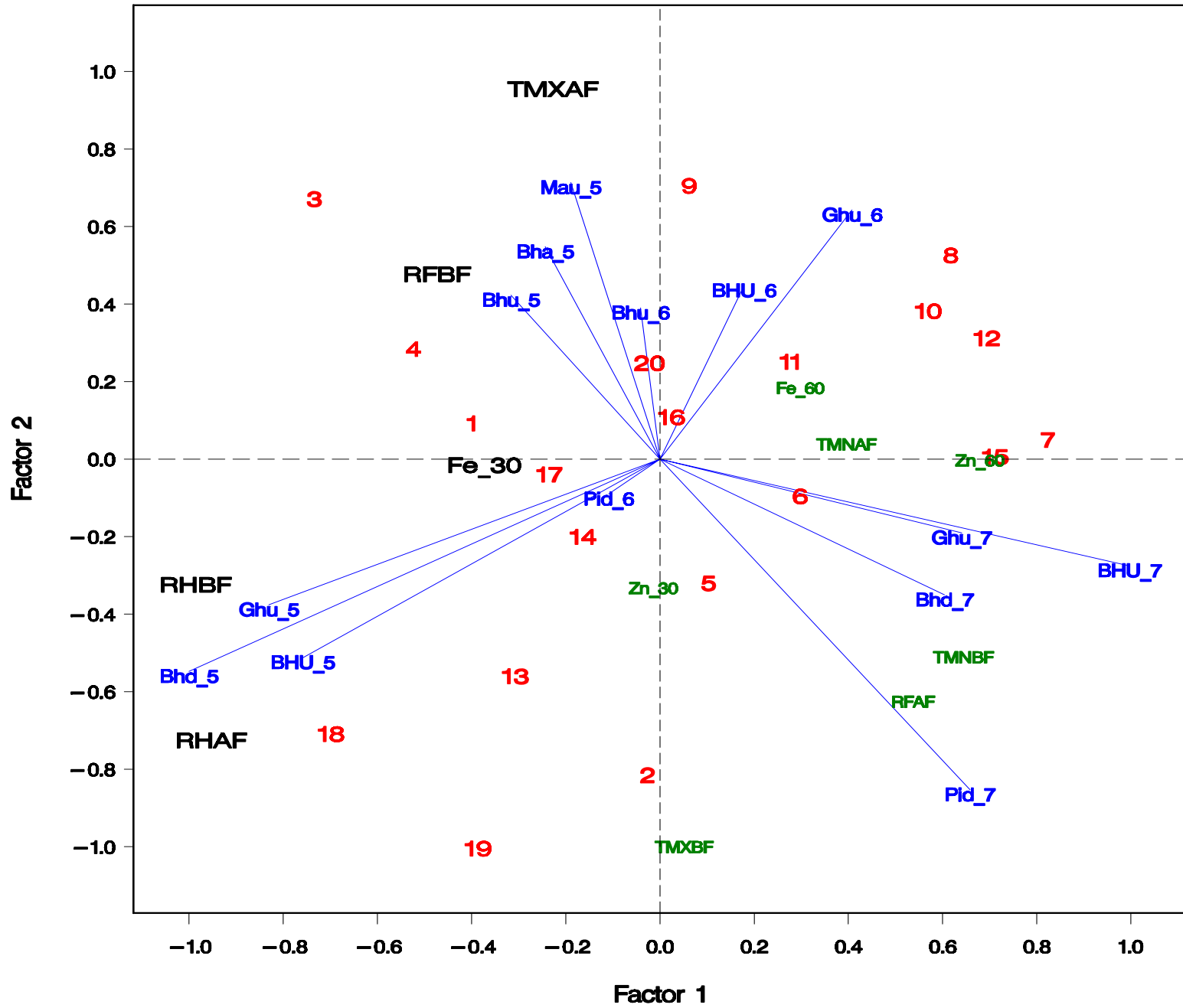
Variable	SS	% variation explained
Grain yield		
RHAF	7671294	13.01
TMXAF	7323023	12.42
TMNBF	7038254	11.94
RHBF	5386681	9.13
Fe_30	4522150	7.67
% contribution of 6 variables		60.26
Fe concentration in the grain		
TMXBF	1045.64	22.21
RFAF	644.96	13.69
Zn_60	644.89	13.69
RHAF	463.93	9.85
% contribution of 4 variables		59.46
Zn concentration in the grain		
TMNAF	1369.06	29.11
Zn_60	751.67	15.98
RFAF	653.90	13.90
TMNBF	604.58	12.85
Zn_30	495.65	10.54
% contribution of 5 variables		82.41

PLS Biplot of number of locations and years with environmental and soil covariables on the performance of grain yield, Fe, and Zn in environments in eastern Gangetic plains of south Asia

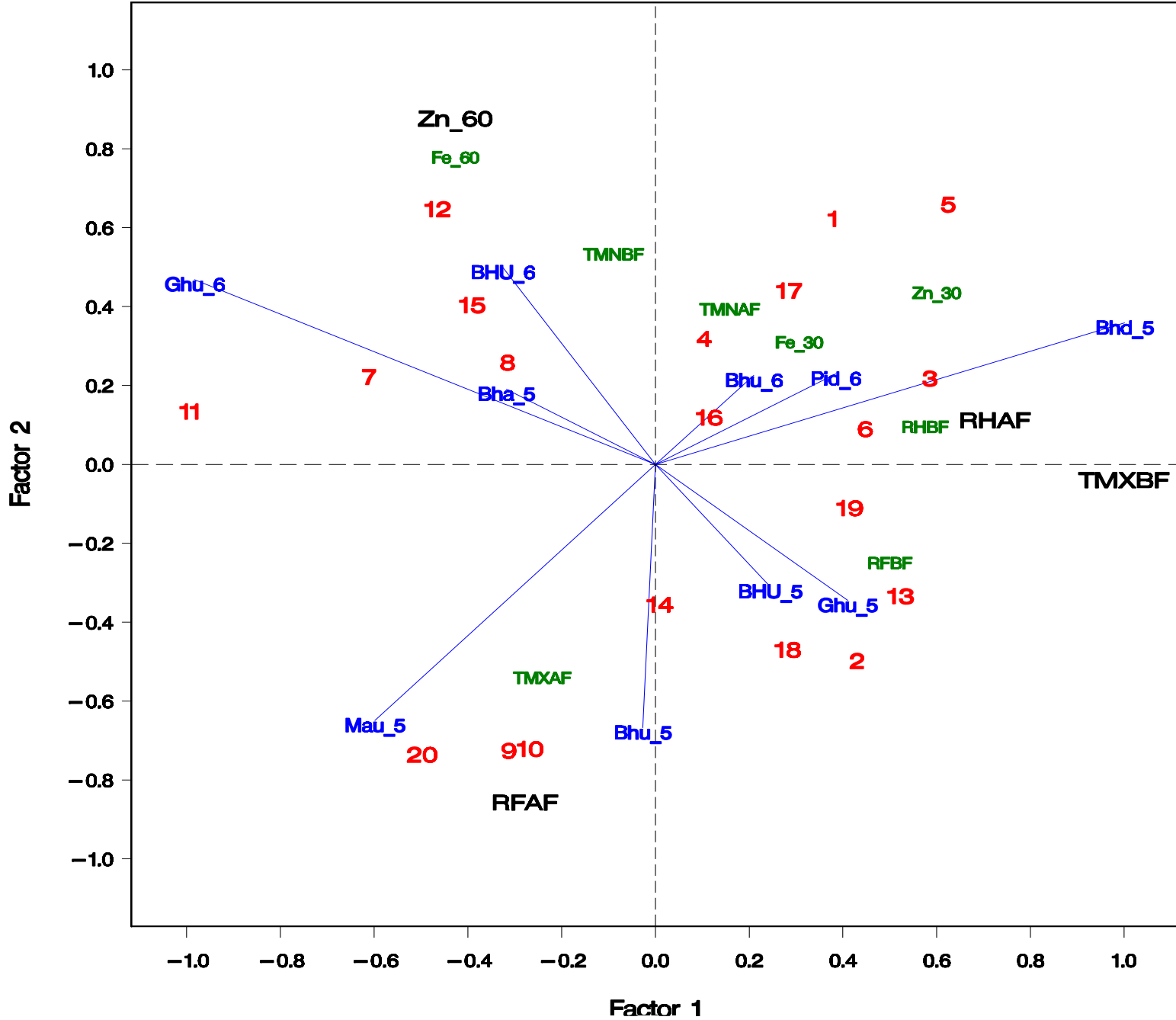
Significant variables are given in bold letters

- ◆ Bhur5= site Bhurkura for year 5; Ghur 5 = Ghurahoopur for year 5; Bhad5 = site Bhadawal for year 5; Bhagwanpur5 = site Bhagwanpur for year 5; Mau5 = site Mauparasi for year 5; BHU5 = site Banaras Hindu University for year 5; Bhur6 = site Bhurkura for year 6; Mau6 = site Mauparasi for year 6; BHU6 = site Banaras Hindu University for year 6; Pidk6 = site Pidkhir for year 6; Bhur7 = site Bhurkura for year 7; Ghur7 = Ghurahoopur for year 7; Pidk7 = site Pidkhir for year 7; BHU7 = site Banaras Hindu University for year 7.
- ◆ TMXBF=Temperature maximum before flowering; TMXAF = Temperature maximum after flowering; TMNBF = Temperature minimum before flowering; TMNAF = Temperature minimum after flowering; RHBF = Relative humidity before flowering; RHAF = Relative humidity after flowering; RBF = Rainfall before flowering (RBF); RAF = Rainfall after flowering (RAF).
- ◆ Zn₃₀ = Zinc concentration in 0-30 cm soil depth; Zn₆₀ = Zinc concentration in 30-60 cm soil depth; Fe₃₀ = Iron concentration in 0-30 cm soil depth; Fe₆₀ = Iron concentration in 30-60 cm soil depth.

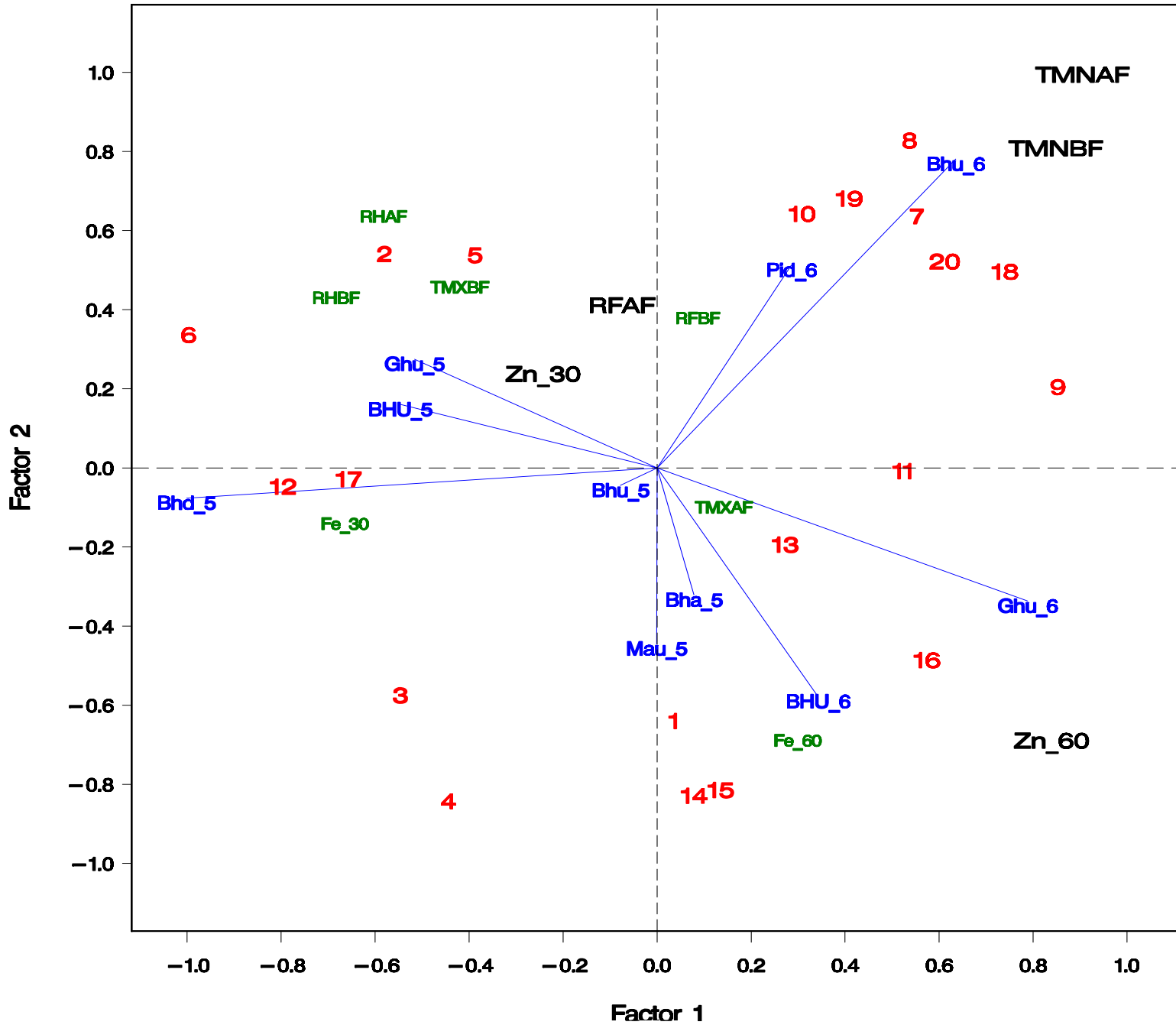
SAMNYT 3 years Yield



SAMNYT 2 years Fe



SAMNYT 2 years Zn

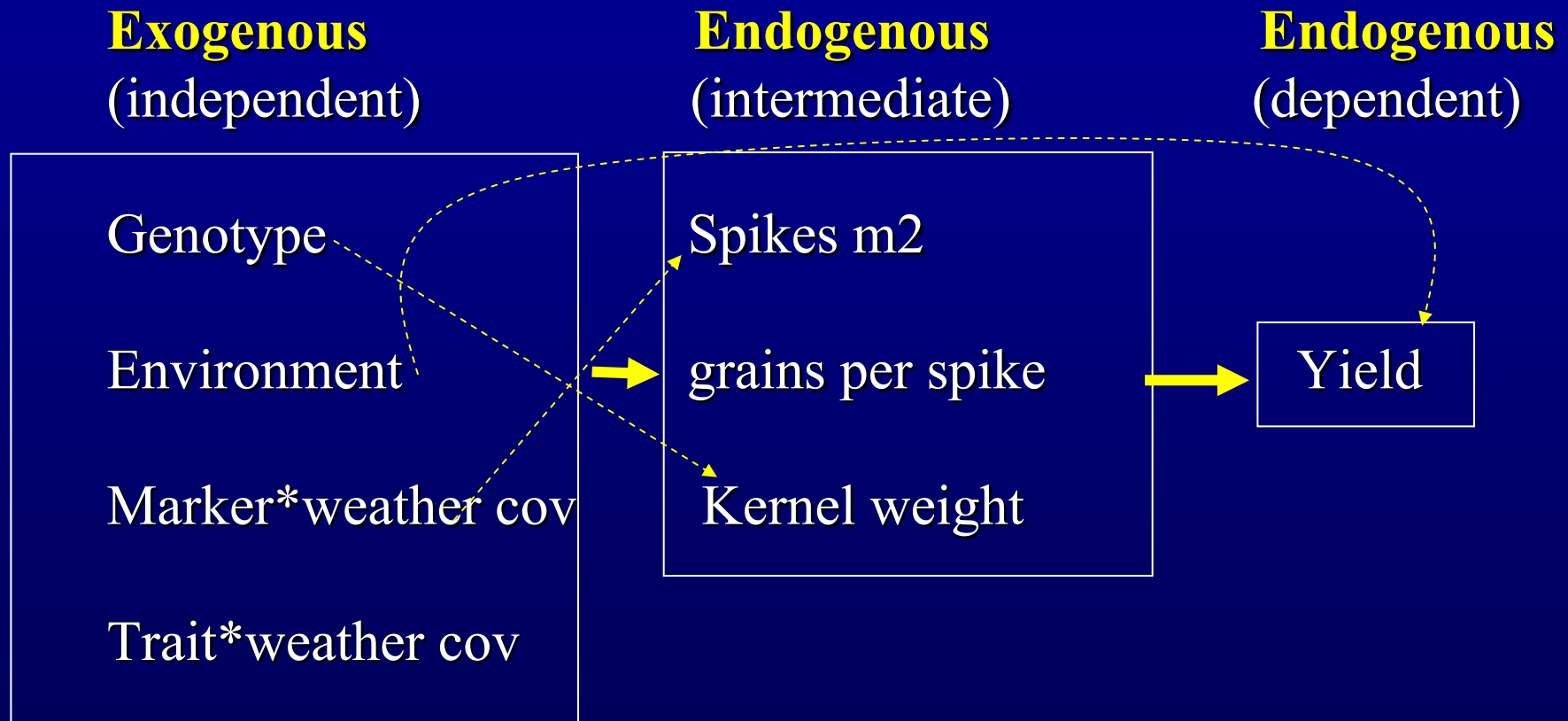


STRUCTURAL EQUATION MODELING (SEM) FOR STUDYING GENOTYPE×ENVIRONMENT

Problems

- ◆ Development of YIELD consists of a sequence of processes
- ◆ Single equation approaches (Factorial Regression or Partial Least Square) are not adequate to incorporate underlying sequential nature of development

Example



Structural equation modeling

SEM

- ◆ Considers several dependent variables at once.
- ◆ Estimates parameters from a system of equations
- ◆ Consider relationship cause-effect

SEM represent the convergence of

- Latent variables models (*psychometric*)
- Simultaneous directional influence (*econometric*)
- Simultaneous linear equations
- Path analysis (*biometric*)

Objective

- ◆ Utilize SEM to explain GE in a complex system of **endogenous** physiological variables and **exogenous** environmental variables.
- ◆ Perform SEM in GE data as well as raw data.

General Structure Equation Models

Latent variables

Interest in modeling variation and covariation of attributes that can not be measured directly (used in social science)

Latent variables are often theoretical concepts (i.e., intelligence, social classes)

→ make measurements on observe variables that are assumed to be indicators of the latent variables

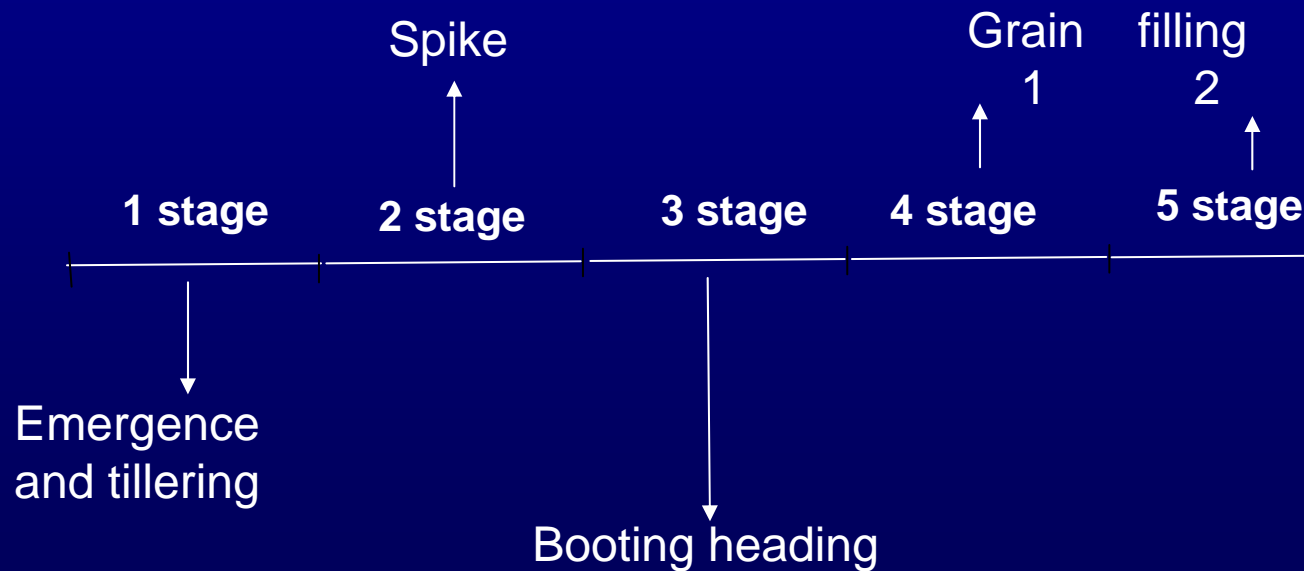
DATA

EXOGENOUS VARIABLES (X)

Climatic covariates x attributes

Weather covariates

mean maximum and minimum temperature, sun radiation



Five stages

DATA

- ◆ 86 wheat genotypes tested in three years.
Incomplete Block design with two replicates.
- ◆ ENDOGENOUS DEPENDENT VARIABLE (Y)
→Yield
- ◆ INTERMEDIATE ENDOGENOUS VARIABLES (Y)
 - 4→ Grains m² (GM2)
 - 4→ Grains per spike (GSP)
 - 4→ Thousand kernel weight (TKW)
 - 4→ Spike m² (SM2)
 - 3→ Biomass anthesis (BMA)
 - 3→ Harvest index anthesis (HIA)
 - 2→ Rapid spike growth (RSG)
 - 2→ Crop growth rate in the boot stage (dBMboot)
 - 1→ Biomass vegetative (BMV)

Fig. 1 Theoretical path diagram

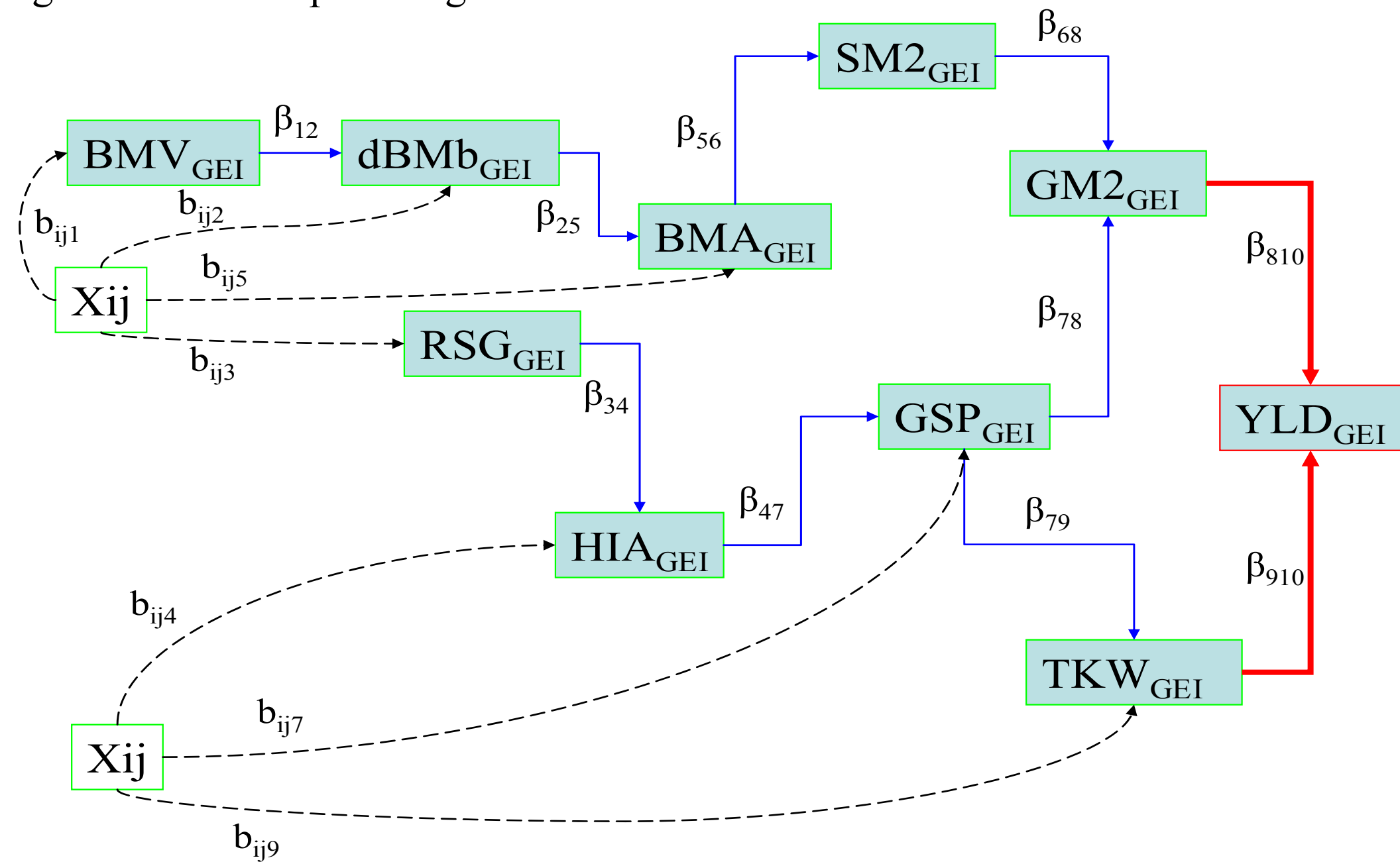
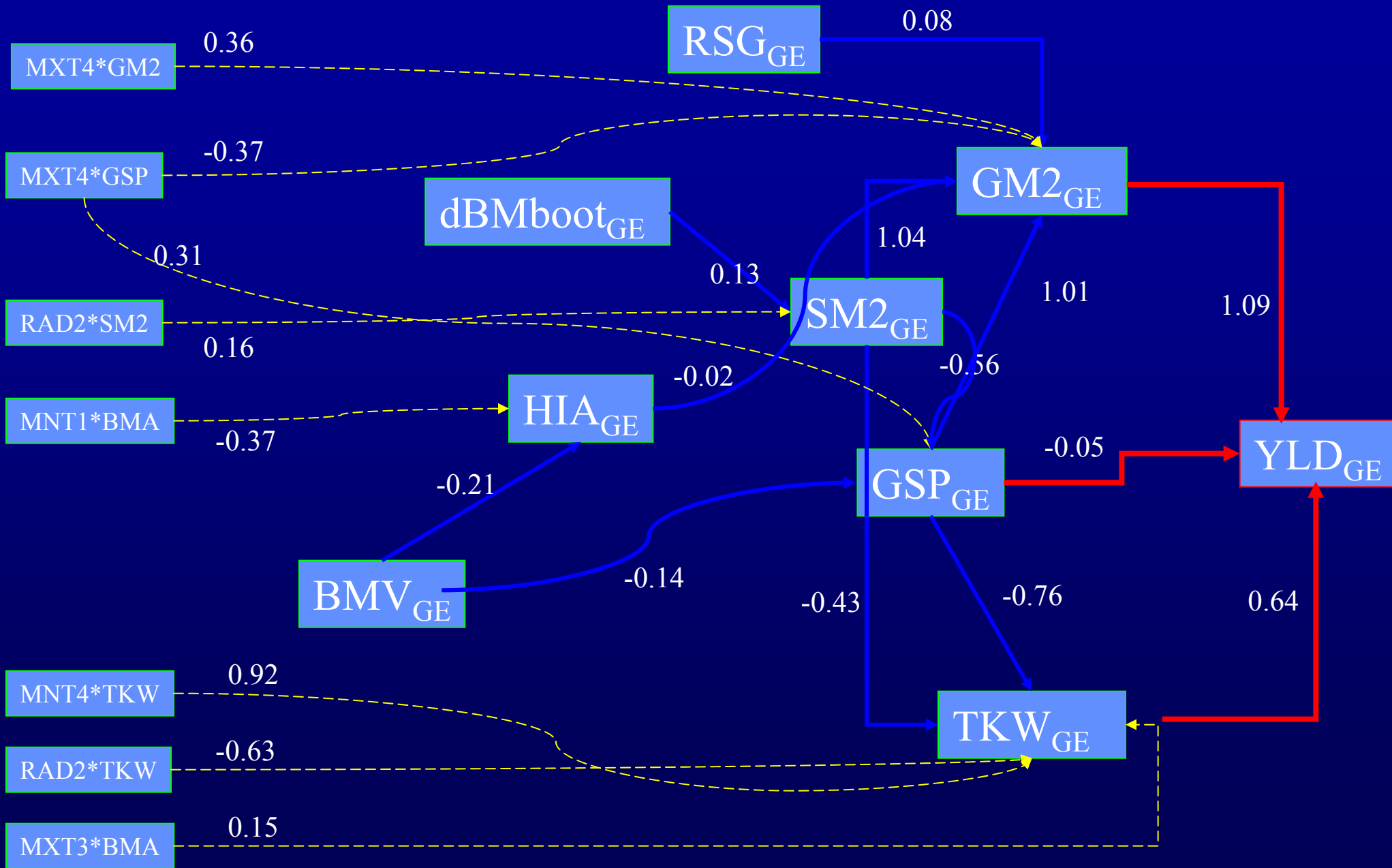


Table 1. Direct, indirect, total, and adjusted cross products effect on YIELD GE ($R^2=0.96$)

Variable	Direct	Indirect	Total	R^2
GM2	1.09	0.00	1.09	0.90
TKW	0.64	0.00	0.64	0.43
GSP	-0.05	0.61	0.56	0.44
SM2	0.00	0.54	0.54	0.04
HIA	0.00	-0.02	-0.02	0.20
RSG	0.00	0.09	0.09	---
DBM _b	0.00	0.07	0.07	---
BMV	0.00	-0.07	-0.07	---
MXT4×GM2	0.00	0.39	0.39	---
MXT4×GSP	0.00	-0.23	-0.23	---
RAD2×SM2	0.00	0.09	0.09	---
MNT4×TKW	0.00	0.59	0.59	---
RAD2×TKW	0.00	-0.40	-0.40	---
MXT3×BMA	0.00	0.10	0.10	---
MNT1×BMA	0.00	0.01	0.01	---

Fig. 2 YLD GE



Summary...

- ◆ FR and PLS are useful tools for incorporating external variables.
- ◆ New QTL and QTL x Environments linear mixed model methods for multi trait multi environment are based on FR.
- ◆ PLS is a shrinkage regression method which is very robust to large number of variables with collinearity.

Summary...

Advantages of SEM over single equation model

- ◆ Provides a comprehensive view of how attributes of the plant and environments work together in a system, using a single path diagram.
- ◆ Provides insight on how GE compensation effects occurs among yield components.
- ◆ Decomposes total effects on GE into direct and indirect effects

Summary...

SEM can be useful for understanding interaction in other applications

- ◆ Complex agronomic interactions

e.g., NITROGEN x WATER on yield through yield components.

- ◆ GEI in human diseases

e.g., gene x smoking on cardiac health as affected through intermediate attributes (blood pressure, arteriosclerosis, etc.)

Biometrics and Statistics Unit

CIMMYT

Mateo Vargas

Juan Burgueno

Jesus Ceron

Gregorio Alvarado

Any Questions
from Guelph,
Ontario,
Canada?

Thank you!!!

