Statistical methods for the analysis of series of plant breeding experiments

Fred van Eeuwijk

Hohenheim, 8-10 october 2007

WAGENINGEN UR
For quality of life

Series of trials = Multi-environment trials

- Single-environment trials typically contain only one treatment *factor of interest: genotype*
- Multi-environment trials typically contain evaluations of sets of genotypes across years and locations, allowing the study of *genotype by environment interactions*
- The trials are usually chosen such as to form a sample from the *target population of environments*
- Sometimes environments are chosen to represent a *known contrast in environmental conditions* (drought versus irrigated)

Some objectives of multi-environment trials

- Study adaptability and stability of genotypes = characterize genotypes
- Characterize environments
- Make an inventory of the genetic and environmental sources of variation shaping phenotypic responses
- Produce global variety predictions
- Produce regional/local variety predictions

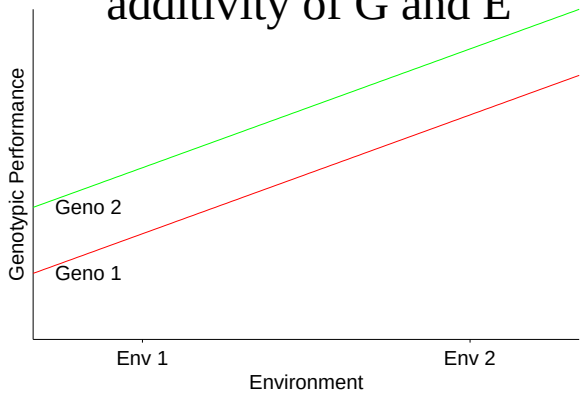
$$P = \int f(\, g(\mathbf{G}) \,,\, e(\mathbf{E}) \,)\, dt \;;\; var(P)=\Sigma$$

- Phenotype is the cumulative result of causal interactions between genetic make-up of plant and environment over developmental time
- Genotype: set of active genes across a (developmental) time interval
- Environment: set of physical factors outside the plant that affects the phenotype by causal interaction with the genetic make-up

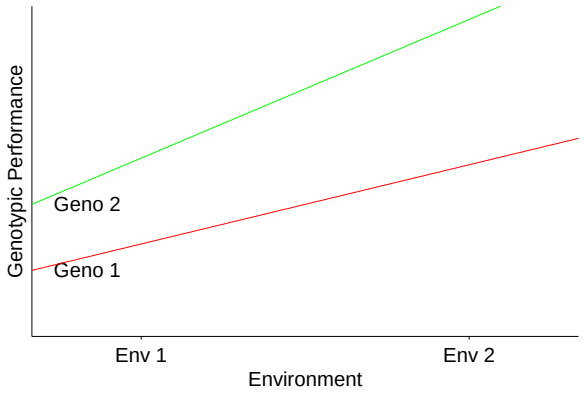
- Genotype by environment interaction: Genotypic differences between phenotypic responses are dependent on the environment
  - GxE for mean:
    - non parallelism of responses
  - GxE for variances and correlations:
    - heterogeneity of variance across environments
    - changing genetic correlations between environments

GxE in terms of changing mean performance across environments

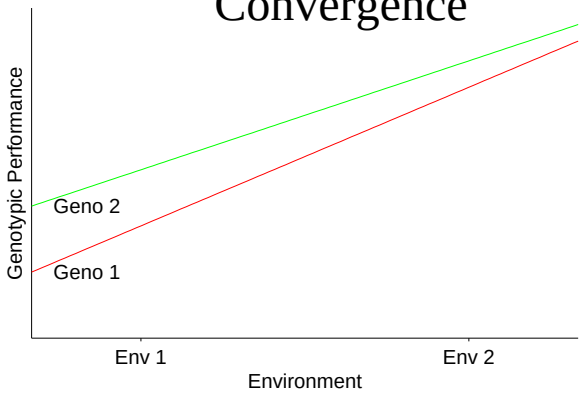
No interaction = additivity of G and E



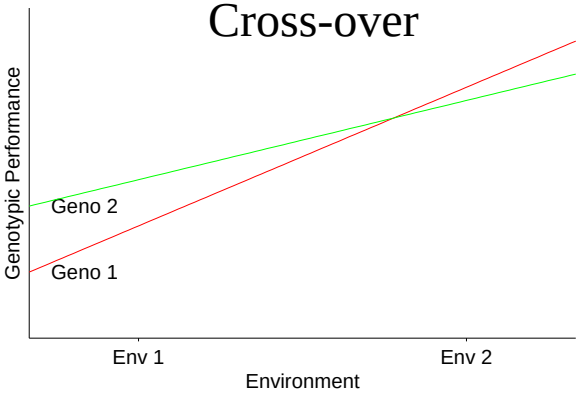
Divergence



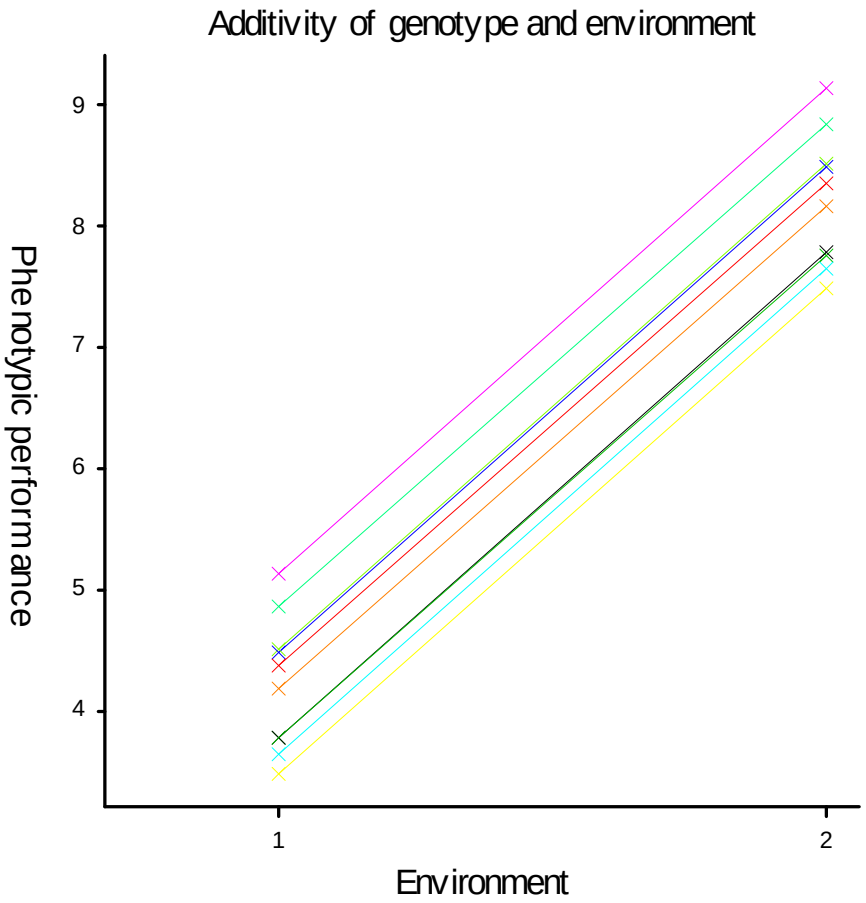
Convergence

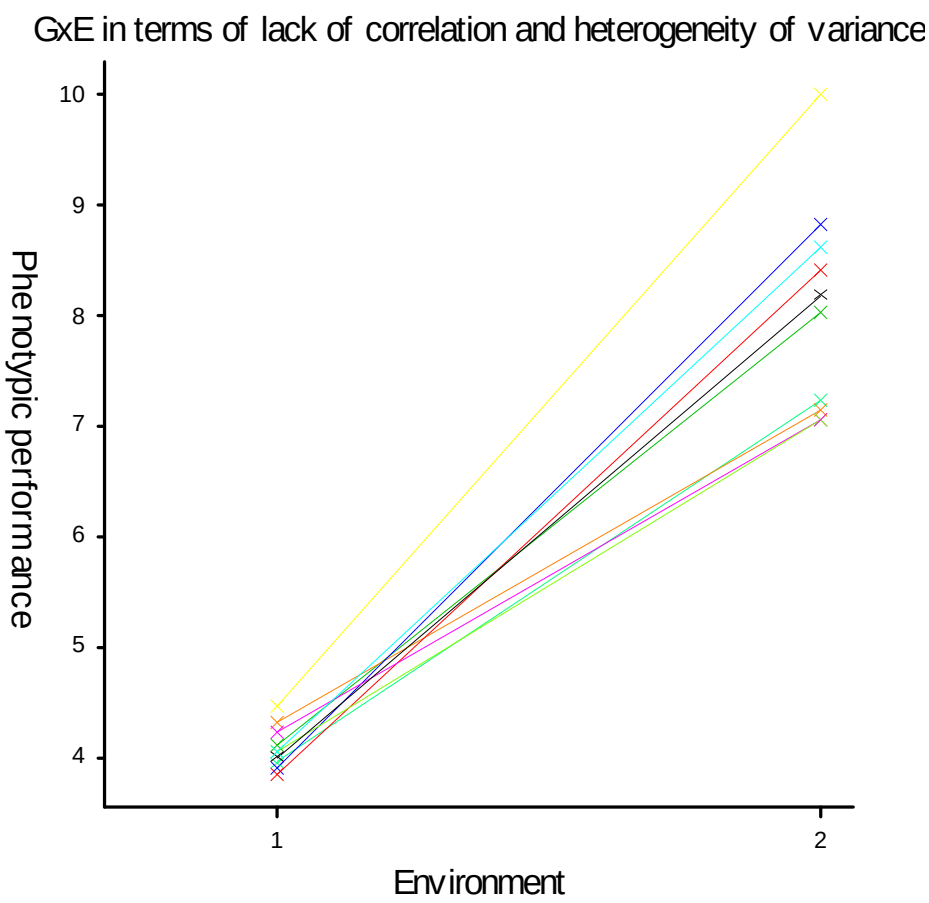


Cross-over

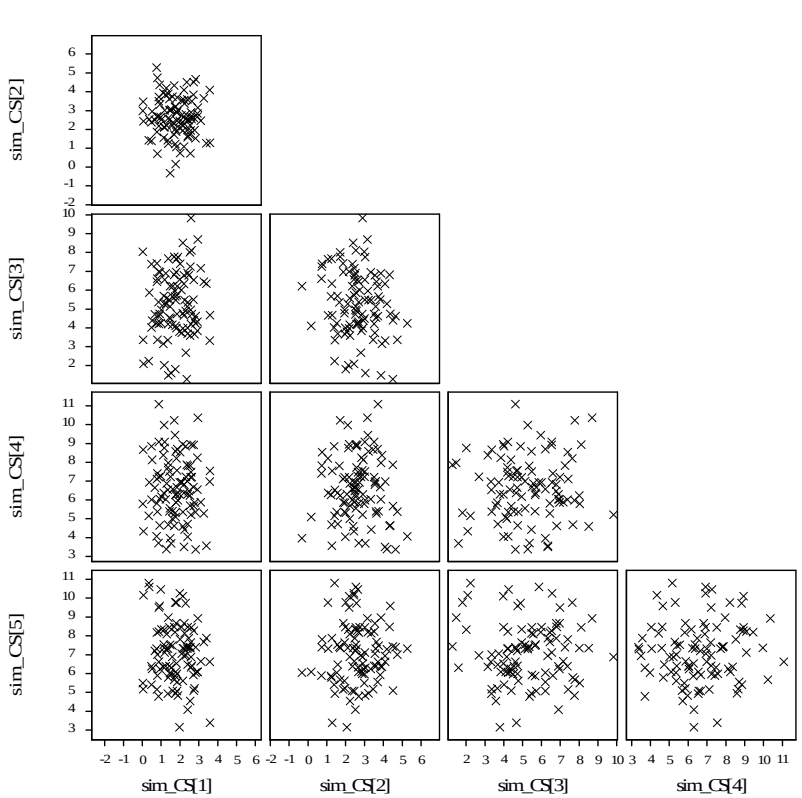


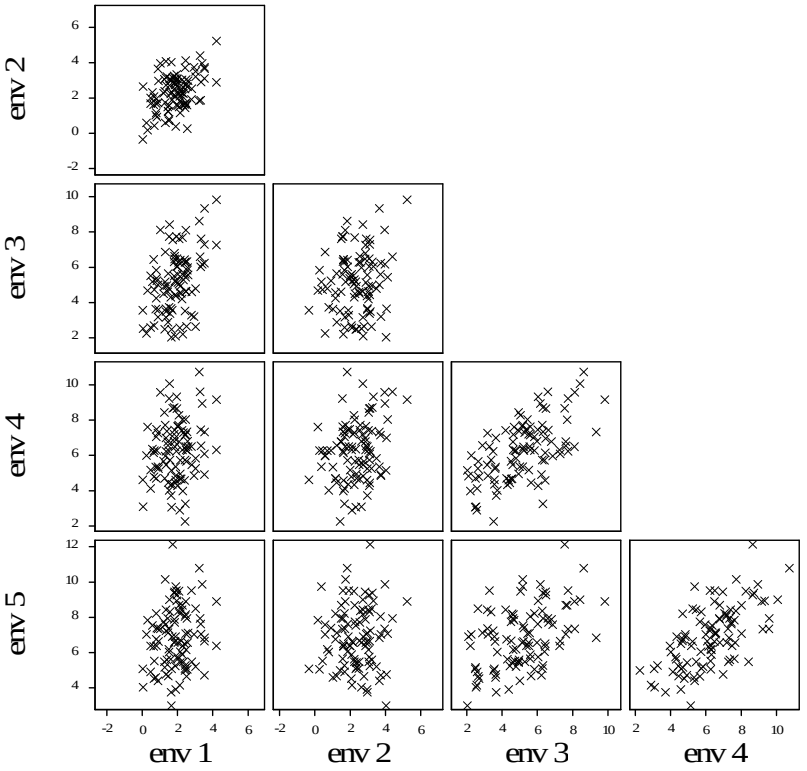
GxE in terms of lack of correlation and heterogeneity of variance





Diagonal (left) versus unstructured (right) VCOV





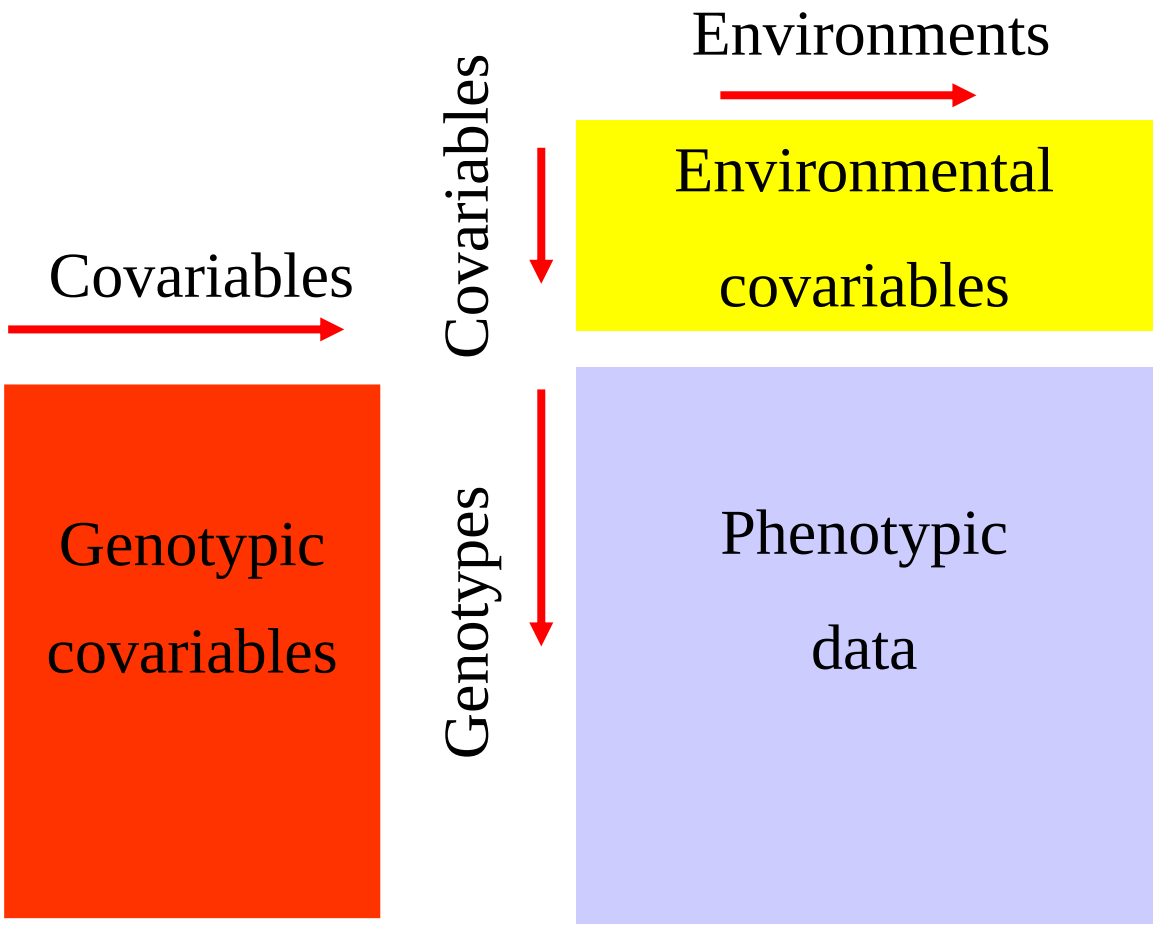
Scatter plot matrix for 5 environments

Modeling mean and VCOV for GxE data

- $\underline{P}_{ij} = \mu_{ij} + \underline{\varepsilon}_{ij}$
  i for genotypes, j for environments

- Aim of statistical modeling for GxE data
  - $\mu_{ij} = f(\boldsymbol{\eta}_i, \boldsymbol{\xi}_j)$ (predictable/ repeatable)
    - Describe μ_{ij} as much as possible in terms of single indexed parameters
    - Find a limited set of double indexed parameters
  - VCOV($\underline{\varepsilon}_{ij}$) (unpredictable/ non-repeatable)
    - Find an appropriate structure for $\underline{\varepsilon}_{ij}$ reflecting heterogeneity of genetic variances and correlations and allowing reliable conclusions on μ_{ij}

Overview of data structures involved in modeling GxE



Models for mean μ_{ij}

Additive model (typical choice var($\underline{\varepsilon}_{ij}$) = σ^2)

$$\mu_{ij} = \mu + E_j + G_i$$

Full interaction model

$$\mu_{ij} = \mu + E_j + G_i + (GE)_{ij}$$

Multiplicative models for interaction

Exploration

Bilinear model $[G_i+]\ (GE)_{ij} = u_{1i}v_{1j} + u_{2i}v_{2j}$

Confirmation

Factorial regression $[G_i+]\ (GE)_{ij} = \kappa x_i z_j / + x_i \alpha_j / + \beta_i z_j$

Modeling GxE: bilinear models (AMMI) and factorial regression

- *AMMI*: $(\underline{GE})_{ij} = u_{1i}v_{1j} + u_{2i}v_{2j} + \underline{\delta}_{ij}$
- v_{1j} and v_{2j} represent theoretical environmental variables that best explain GxE through the genotypic sensivities u_{1i} and u_{2i}.
- Biplot: graphical representation of AMMI model for exploratory purposes of GxE

  = Plot genotypes as (u_{1i}, u_{2i}) and environments as (v_{1j}, v_{2j})

- *Factorial regression*: $(\underline{GE})_{ij} = \kappa x_i z_j + x_i \alpha_j + \beta_i z_j + \underline{\delta}_{ij}$
- x_i = genotypic covariable, z_j = environmental covariable
- α_j = environment related characterization
- β_i = genotypic sensitivity
- κ = scaling constant with respect to genotype by environment characterizations $(GE)_{ij} = \kappa x_i z_j = \kappa(xz)_{ij}$

VCOV models for GxE data

$$\underline{P}_{ij} = \mu_{ij} + \underline{\varepsilon}_{ij}$$

$\text{var}(\varepsilon_{ij})$	$cov(\varepsilon_{ij};\varepsilon_{ij*}),\ j \neq j*$	n_{PAR}	Description
σ^2_G	0	1	No cov., homog. var.
$\sigma^2_G + \sigma^2_{GE}$	σ^2_G	2	Uniform cov., homog. var.
$\sigma^2_{G_j}$	0	J	No cov., heterog. var.
$\sigma^2_G + \sigma^2_{GE_j}$	σ^2_G	J+1	Uniform cov., heterog. var.
$\sigma^2_{G_j}$	$\theta\sigma_{G_j}\sigma_{G_{j*}}$	J+1	Uniform corr., heterogen. var.
$\lambda^2_j + \sigma^2_{d_j}$	$\lambda_j\lambda_{j*}$	2J	Factor analytic
$\sigma^2_{G_j}$	$\sigma_{G\,jj*}$	J(J+1)/2	Unstructured

VCOV Diagonal on environments

$$\mu_{ij} = \mu + E_j$$

$$\mu_{ij} = \mu + E_j + x_i\alpha_j$$

$$\mu_{ij} = \mu + E_j + G_i$$

$$\mu_{ij} = \mu + E_j + G_i + \beta_i z_j$$

$$VCOV(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \sigma^2_1 & & & \\ & \sigma^2_2 & & \\ & & \sigma^2_3 & \\ & & & \sigma^2_4 \end{bmatrix}$$

$$Corr(Env_j; Env_{j^*}) = \frac{0}{\sigma_j \sigma_{j^*}} = 0$$

Each environment has its own (residual) genetic variance (that is confounded with GxE variance). There is no genetic correlation between environments

VCOV Compound symmetry / Uniform correlation on Env

$$\mu_{ij} = \mu + E_j$$

$$\mu_{ij} = \mu + E_j + x_i \alpha_j$$

$$VCOV(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \sigma_G^2 + \sigma_{GE}^2 & & & \\ \sigma_G^2 & \sigma_G^2 + \sigma_{GE}^2 & & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_{GE}^2 & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_{GE}^2 \end{bmatrix}$$

$$Corr(Env_j ; Env_{j^*}) = \frac{\sigma_G^2}{\sqrt{\sigma_G^2 + \sigma_{GE}^2}\sqrt{\sigma_G^2 + \sigma_{GE}^2}} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2}$$

VCOV Unstructured on environments

$$\mu_{ij} = \mu + E_j$$

$$\mu_{ij} = \mu + E_j + x_i\alpha_j$$

$$\mu_{ij} = \mu + E_j + x_i\alpha_j + \beta_i z_j$$

$$VCOV(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

$$Corr(Env_j; Env_{j^*}) = \frac{\sigma_{jj^*}}{\sigma_j \sigma_{j^*}}$$

VCOV: Factor analytic on environments

$$\mu_{ij} = \mu + E_j$$

$$\mu_{ij} = \mu + E_j + x_i\alpha_j$$

$$\mu_{ij} = \mu + E_j + x_i\alpha_j + \beta_i z_j$$

Heterogeneity of variances and correlations possible at the price of relatively few parameters

This kind of structure allows reliable (not too optimistic) tests for QTL main effects and QTLxE

$$VCOV(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \lambda_1\lambda_1 + \delta_1^2 & & & \\ \lambda_2\lambda_1 & \lambda_2\lambda_2 + \delta_2^2 & & \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3\lambda_3 + \delta_3^2 & \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4\lambda_4 + \delta_4^2 \end{bmatrix}$$

$$Corr(Env_j; Env_{j^*}) = \frac{\lambda_j\lambda_{j^*}}{\sqrt{(\lambda_j\lambda_j + \delta_j^2)(\lambda_{j^*}\lambda_{j^*} + \delta_{j^*}^2)}}$$

Inference

- Standard linear mixed model framework is available
  - Wald tests for fixed effects
  - Deviance tests for VCOV parameters
- AIC and BIC for non nested models
- Multiple test corrections in genotypic and environmental covariable selection
  - Bonferroni
  - False discovery rates

GxE data: Example phenotypic analyses

Ignacio Romagosa, Fred van Eeuwijk & Bill Thomas
Mapping Adaptation of Barley to Droughted Environments
EU FP5 INCO-MED ICA3-CT2002-10026

WAGENINGEN UR
For quality of life

Mapping adaptation to drought stress in barley
65 genotypes in 12 environments

Location	Code	Latitude	Longitude	Altitude (m)	Sowing date	Yield (t/ha)	Intrablock $\hat{\sigma}^2_{error}$	H^2
El Khroub, Algeria	A4	36º32'N	06ª42'E	596	05/12/2003	5.29	0.76	0.38
El Khroub, Algeria	A5	36º32'N	06ª42'E	596	19/02/2005	3.26	0.33	0.49
Gimenells, Spain	E4	41º35'N	00º32'E	260	17/12/2003	6.79	0.13	0.54
Gimenells, Spain	E5	41º35'N	00º32'E	260	24/11/2004	3.05	0.05	0.53
Foggia, Italy	I4	41º27'N	15º34'E	57	13/01/2004	4.77	0.10	0.83
Foggia, Italy	I5	41º27'N	15º34'E	57	16/12/2004	5.26	0.19	0.53
Settat, Morocco	M4	33º07'N	07º37'W	240	16/12/2003	4.23	0.66	0.63
Settat, Morocco	M5	33º07'N	07º37'W	240	11/12/2004	1.61	0.20	0.50
Tel Hadya, Syria	S4	36º01N	36º56'E	362	11/12/2003	3.97	0.24	0.09
Tel Hadya, Syria	S5	36º01N	36º56'E	362	01/12/2004	5.08	0.22	0.54
Haymana, Turkey	T4	39°36'N	32°40'E	1214	01/03/2004	6.53	0.44	0.54
Esenboga, Turkey	T5	40°08'N	33°01'E	953	21/03/2005	4.88	0.43	0.53
					Median	4.82	0.23	0.53

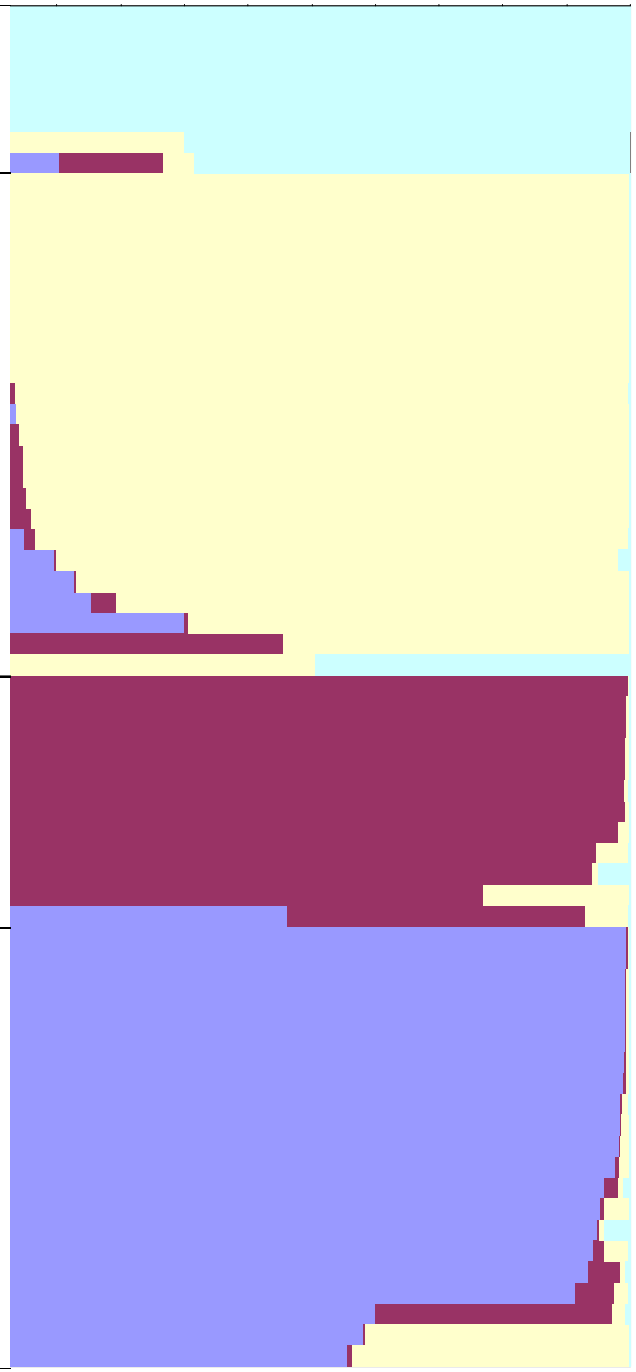
Turkish lines

North Mediterranean Winter (2-6 rows)

South West Mediterranean (6 rows)

North Mediterranean (2 rows)

	Entryname	Origin	Spike type	Phenology
M19	Orza96	TUR	2	S
M14	Anadolu98	TUR	2	S
M17	Karatay	TUR	2	S
M18	Tarm92	TUR	2	S
M13	Efes98	TUR	2	S
M21	Yesevi93	TUR	2	S
M16	Sahin91	TUR	2	W
M20	Bulbul89	TUR	2	S
M33	Aliseo	ITA	6	W
M48	Solen	ITA	6	W
M38	Federal	ITA	6	W
M07	Manitou	FRA	6	W
M12	Intro	NLO	6	W
M40	Grecale	ITA	2	W
M42	Kelibia	ITA	2	W
M49	Sonora	ITA	6	W
M45	Nure	ITA	2	W
M34	Amillis	ITA	2	W
M06	Igri	DEU	2	W
M53	Ultra	ITA	2	W
M39	Gotic	ITA	6	W
M44	Mattina	ITA	6	W
M30	Reinette	FRA	2	W
M10	Siberia	FRA	6	W
M65	Manel	DZA	6	S
M54	Vertige	ITA	2	W
M05	Fanfare	GBR	2	W
M25	Hispanic	FRA	2	W
M46	Naturel	ITA	2	W
M23	Dobla	ESP	6	S
M22	Candela	ESP	6	S
M15	**Aydanhanin**	**TUR**	**2**	**W**
M60	Oussama	MAR	6	S
M58	Massine	MAR	2	S
M59	Rabat01	MAR	6	S
M01	Arig8	MAR	6	S
M57	Merzaga07	MAR	6	S
M61	ASCAD 176	JOR	6	S
M31	Steptoe	USA	6	S
M29	Orria	ESP	6	S
M64	Alanda01	DZA	6	S
M62	Rum	JOR	2	S
M56	**Amalou**	**MAR**	**6**	**S**
M55	**Aglou**	**MAR**	**2**	**S**
M35	Apex	ITA	2	S
M50	Tea	ITA	2	S
M36	Barke	ITA	2	S
M41	Grosso	ITA	2	S
M11	Triumph	DEU	2	S
M43	Magda	ITA	2	S
M28	Nevada	GBR	2	S
M04	Chariot	GBR	2	S
M02	Atem	NLO	2	S
M47	Otis	ITA	2	S
M03	Alexis	DEU	2	S
M63	Aramir	NLO	2	S
M09	Scarlett	DEU	2	S
M52	Tremois	ITA	2	S
M32	Zaida	ESP	2	S
M24	Graphic	GBR	2	S
M27	Kym	GBR	2	S
M08	Optic	GBR	2	S
M26	Kika	ESP	2	S
M51	**Tiobne**	**ITA**	**2**	**S**
M37	**Dasio**	**ITA**	**2**	**S**



Characterizing genotypes

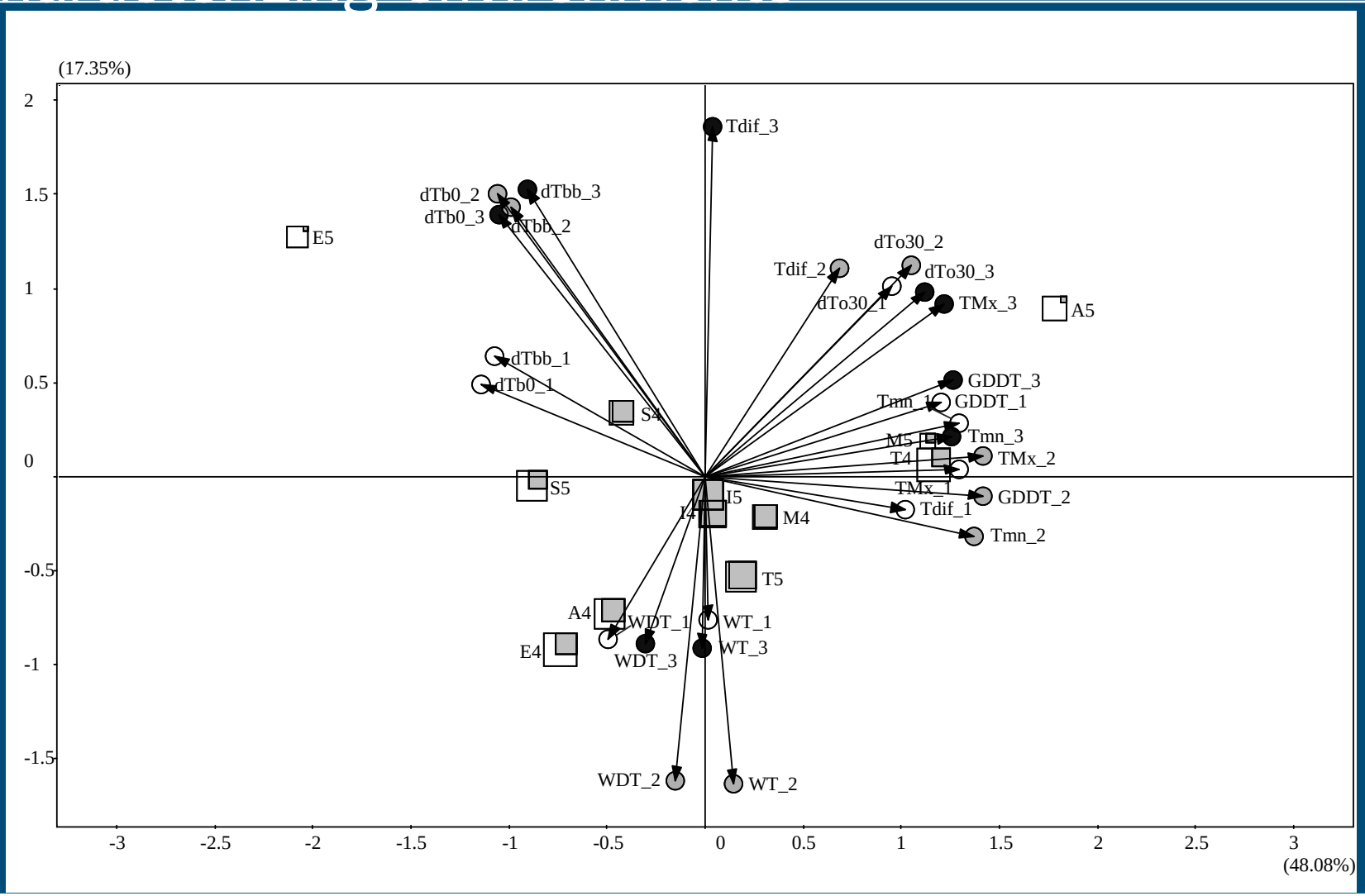
on genotypic information

WAGENINGEN UR
For quality of life

Environmental characterization (tillering, jointing, and grain filling)

- Number of days with minimum temperature below 0 degrees (dTb0)
- Number of days with average daily temperature below 4 degrees (base temperature) (dTbb)
- Number of days with maximum temperature over 30 degrees (dTo30)
- Average maximum temperature (TMx)
- Average minimum temperature (TMn)
- Difference between the maximum and the minimum temperature (Tdif)
- Total Growing Degree Days (GDD)
- Total rainfall plus irrigation (mm) (WT)
- Ratio between available water to Evapotranspirative demand, 100*Total water/Et0, (WDT)
- Average photo thermal quotient, Solar radiation/average daily temperature, (PQ)

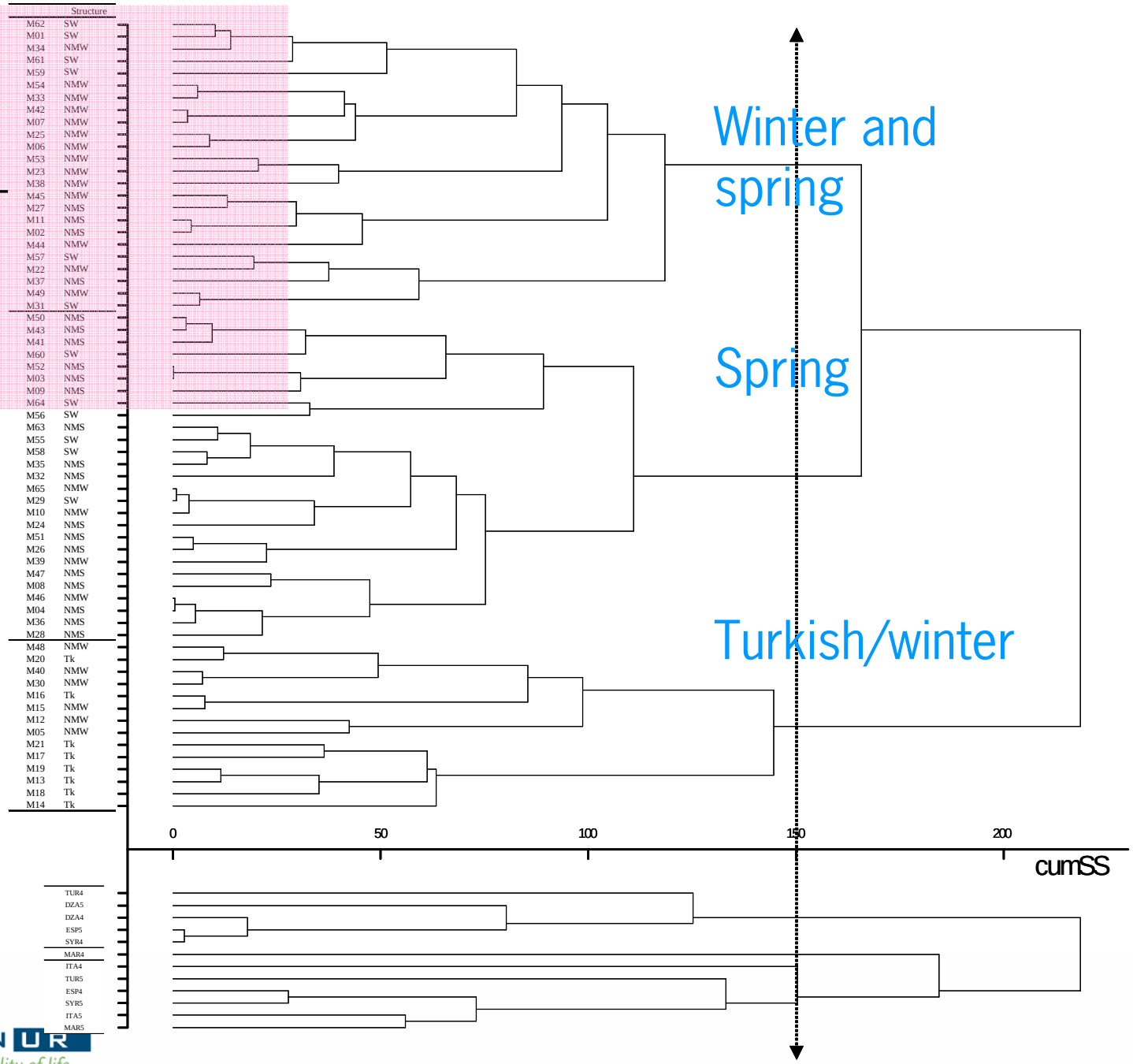
Characterizing environments



$$\overline{P}_{i(k)j(l)} = \mu + [GC_k + G'_{i(k)}] + [EC_l + E'_{j(l)}] + [(GC.EC)_{kl} + (G.E)'_{i(k)\,j(l)}]$$

Characterizing genotypes and environments

on GxE



WAGENINGEN UR
For quality of life

Characterizing genotypes and environments

$$\overline{P}_{i(k)j(l)} = \mu + [GC_k + G'_{i(k)}] + [EC_l + E'_{j(l)}] + [(GC.EC)_{kl} + \underline{(G.E)}'_{i(k)\,j(l)}]$$

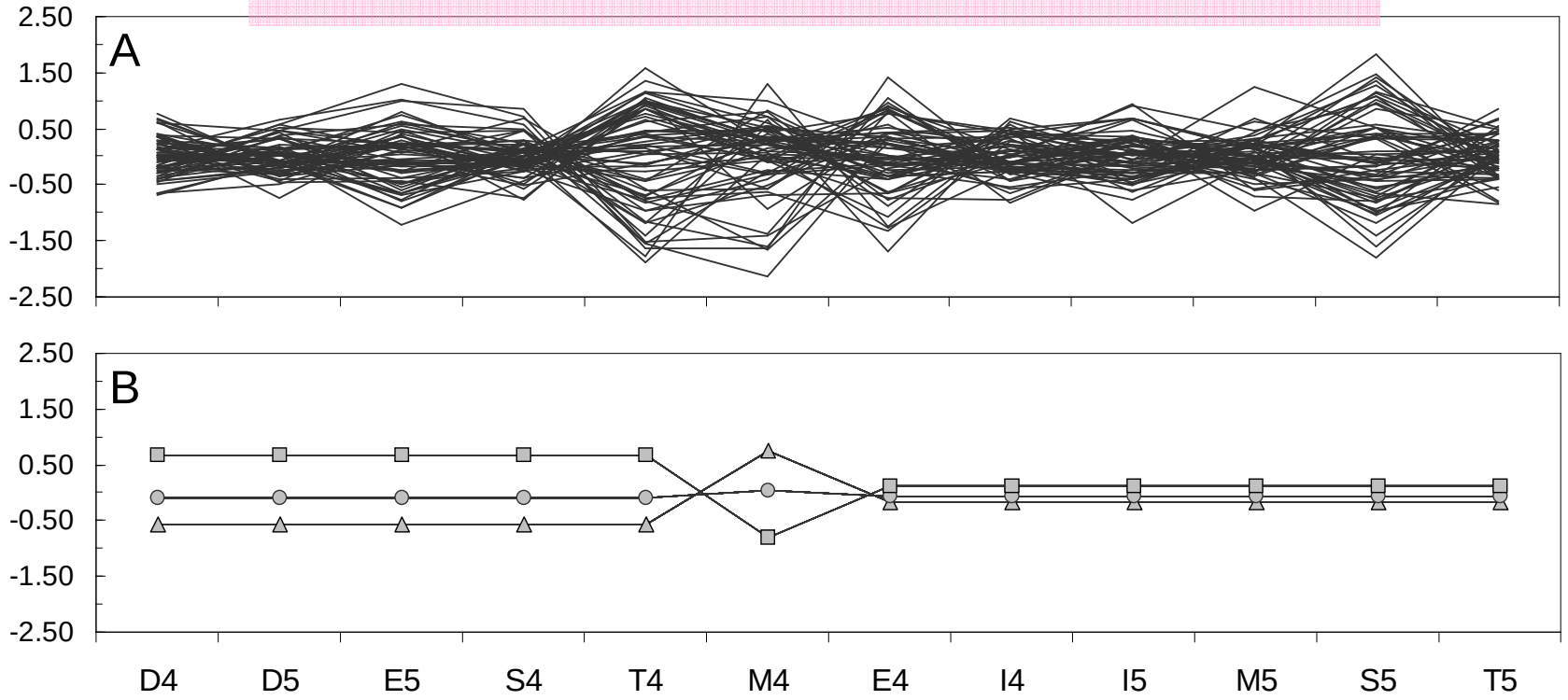
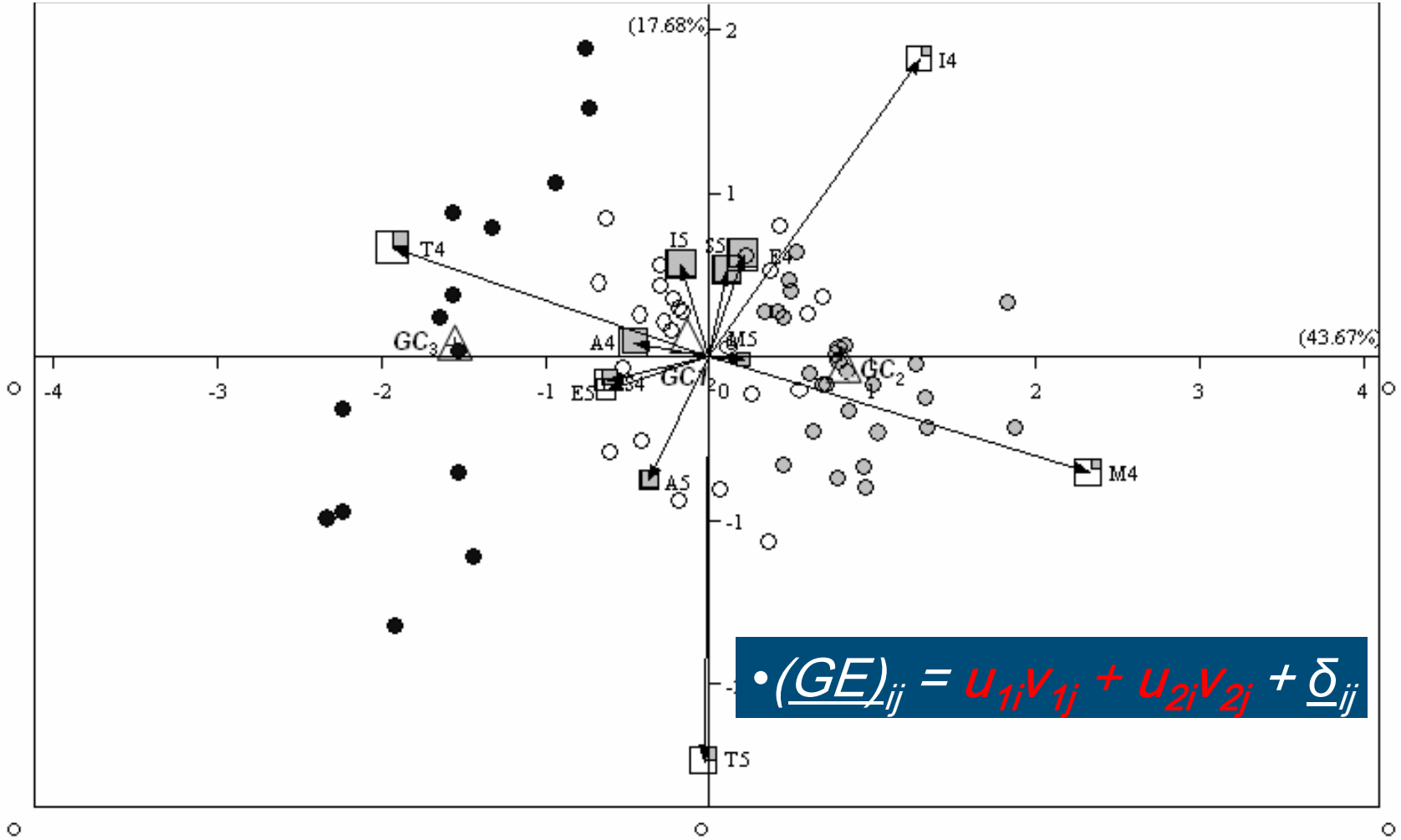


Figure 6. AMMI biplot. Genotypes are represented in circle (open, gray and dark representing the three Corsten & Denis' (1990) clusters identified in the text) The triangles represent the mean for the three clusters. Environments are shown in squares with areas proportional to its average yield. Within each square the solid portion is a representation of the amount of the sum of squares for each environment which is not explained by the axes under consideration.



$$\bullet (\underline{GE})_{ij} = u_{1i}v_{1j} + u_{2i}v_{2j} + \underline{\delta}_{ij}$$

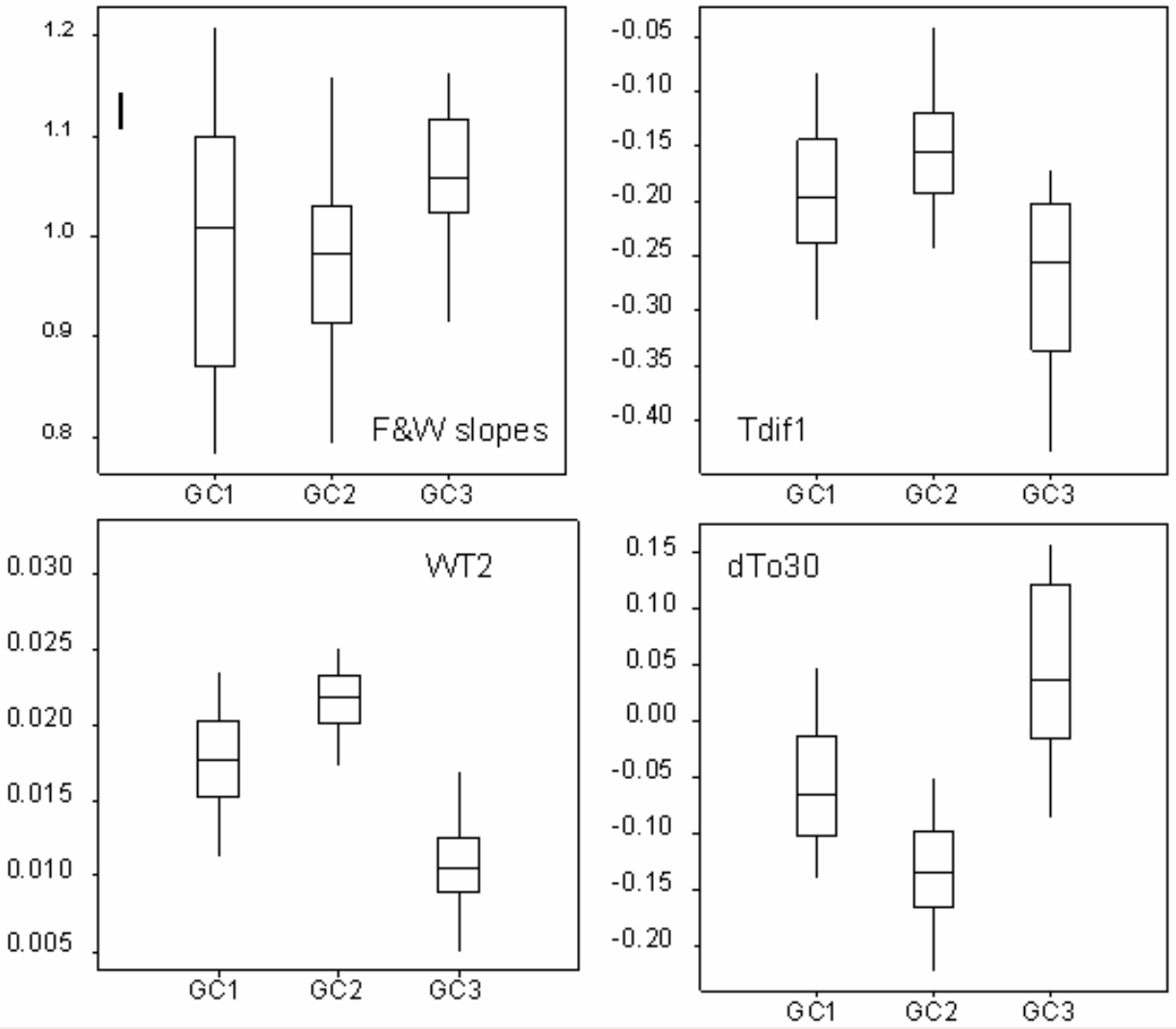
$$\mu_{i(k)j} = \mu + [GC_k + \beta_k z_j] + [G'_{i(k)} + \beta'_{i(k)} z_j]$$

Table 5. Partitioning of the G.E interaction in Table 1 according to a multiple factorial regression based on the average difference between daily maximum and minimum temperature during Tillering (Tdif), Total Water (WT) at Jointing and days with temperature over 30°C (dTo30) at grain filling.

Source of variation	d.f	Sum of squares	Semipartial R^2	Mean squares	variance ratio $\equiv F_c$	$-\log_{10}(p)$
Environment [E]	*11*	*1522.2*	*85.3*	*138.4*	*445.8*	
Tdif (Tillering)	1	138.5	9.1	138.5	446.3	*>100*
WT (Jointing)	1	426.8	28.0	426.8	1375.2	*>100*
dTo30 (Grain Filling)	1	121.0	7.9	121.0	389.8	*>100*
E'	8	835.9	54.9	104.5	336.6	*>100*
Genotype [G]	*64*	*44.4*	*2.5*	*0.69*	*2.23*	
GE	*704*	*218.5*	*12.2*	*0.31*	*0.99*	
Tdif.*GC*	2	6.4	2.9	3.19	11.18	*4.75*
WT.*GC*	2	6.7	3.1	3.37	11.81	*5.02*
dTo30.*GC*	2	28.1	12.9	14.05	49.33	*19.59*
Tdif.Genotype(*GC*')	62	12.0	5.5	0.19	0.68	*0.01*
WT.Genotype(*GC*')	62	9.8	4.5	0.16	0.56	*0.00*
dTo30.Genotype(*GC*')	62	9.6	4.4	0.15	0.54	*0.00*
Residual	512	145.8	66.7	0.28		

$$\mu_{i(k)j} = \mu + [GC_k + \beta_k z_j] + [G'_{i(k)} + \beta'_{i(k)} z_j]$$

Figure 5. Box plots for the slopes of the regression on the means model (F&W Slopes) and the factorial regression derived genetic sensitivities for the Tdif1, WT2 and dTo30 classified according to the genotypic clusters GC_1 to GC_3 (for the acronyms of meteorological variables see text)



1. Winter and spring
2. Spring
3. Turkish/winter

Modeling mean and VCOV for GxE data

- $\underline{P}_{ij} = \mu_{ij} + \underline{\varepsilon}_{ij}$

  i for genotypes, j for environments

- Aim of statistical modeling for GxE data
  - $\mu_{ij} = f(\boldsymbol{\eta}_i; \boldsymbol{\xi}_j)$ (predictable/ repeatable)
    - Describe μ_{ij} as much as possible in terms of single indexed parameters
    - Find a limited set of double indexed parameters
  - VCOV($\underline{\varepsilon}_{ij}$) (unpredictable/ non-repeatable)
    - Find an appropriate structure for $\underline{\varepsilon}_{ij}$ reflecting heterogeneity of genetic variances and correlations and allowing reliable conclusions on μ_{ij}

Publications
(for pdf's, marcos.malosetti@wur.nl)

- Malosetti M, Boer MP, Bink MCAM, van Eeuwijk FA (2006) Multi-trait QTL analysis based on mixed models with parsimonious covariance matrices. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, August 13-18, Belo Horizonte, MG, Brasil. http://www.wcgalp8.org.br/wcgalp8/ Article 25-04
- Malosetti M, Ribaut JM, Vargas M, Crossa J, Boer MP, van Eeuwijk FA (2007) Multi-trait multi-environment QTL modelling for drought-stress adaptation in maize. In: JHJ Spiertz, Struik PC,van Laar HH (Eds.), Scale and Complexity in Plant Systems Research. Gene-Plant-Crop Relations, pp. 25-36. Springler, Dordrecht, The Netherlands.
- van Eeuwijk FA, Malosetti M, Boer MP (2007) Modelling the genetic basis of response curves underlying genotype x environment interaction. In: JHJ Spiertz, Struik PC,van Laar HH (Eds.), Scale and Complexity in Plant Systems Research. Gene-Plant-Crop Relations, pp. 115-126. Springler, Dordrecht, The Netherlands.
- Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk FA (2007) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). Euphytica (in press)
- Martin Boer, Deanne Wright, Lizhi Feng, Dean Podlich, Lang Luo, Mark Cooper, Fred van Eeuwijk (2008) A mixed model QTL analysis for multiple environment trial data using environmental covariables for QTLxE, with an example in maize. Genetics (in press).
- Mathews KL, Malosetti M, Chapman SC, McIntyre L, Reynolds MP, Shorter R, van Eeuwijk FA Multi-environment QTL mixed models for drought stress adaptation in wheat. Submitted.