# Modelling GxE-interaction in plants

## A quest for structure

Marcos Malosetti
Daniela Bustos-Korts, Fred van Eeuwijk & Piter Bijma
Wageningen University and Research
Armidale, January 2017

WAGENINGEN UR
For quality of life

# 1. Introduction

WAGENINGEN UR
For quality of life

## Causes of GxE in plants

- Observed GxE (i.e. statistical)
  - Spatial: location-effects (L)
  - Temporal: year-season effects (Y)
  - L*Y-interactions
- Mechanism
  - Rainfall, temperature, soil, daylight, ....
- GxE often more important than in animals
  - Wider range of environments?
  - Immobility

WAGENINGEN UR
For quality of life

## GxE in animal vs plant breeding

- Animal breeding:
  - Multitrait mixed models (Character-state model)
  - Each trait in each environment is a different "trait"
  - Focus on genetic correlation of the same trait in two environments ($r_g$).
  - Very little structure
    - We cannot predict a third environment

- Plant breeding
  - Identify causes of GxE → towards predictability
  - Separate predictable (L) from unpredictable (Y) GxE
  - Separate G and E components of the GxE
  - Structured models

## Typical GxE data in plant breeding

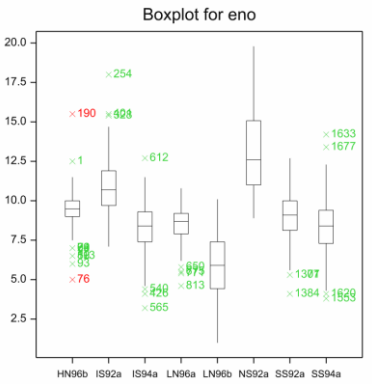
- Multi-Environment Trials (MET):
  - Genotypic dimension:
    - Specific sets of genotypes.
    - A sample from a population of interest.
    - Unstructured vs structured (genetic relationships).
  - Environmental dimension:
    - Target Population of Environments (TPE).
    - Unstructured vs structured set of environments.
      - Environmental variables (°C)

## Typical research questions on GxE

- Related with the genotypes:
  - **Adaptation:** are particular genotypes adapted to certain environmental range?
  - **Adaptability** / sensitivity: are particular genotypes able to become adapted to (changes) improvements in the environment?
  - **Stability:** is the performance of particular genotypes consistent?
- Related with the environments:
  - **Structure:** Grouping of trials into mega-environments: finding structure in the TPE.
  - **Design:** Given a structure of the TPE optimize the choice of trials to represent the TPE.

# Crude indications of GxE: 1. Heterogeneity of variance

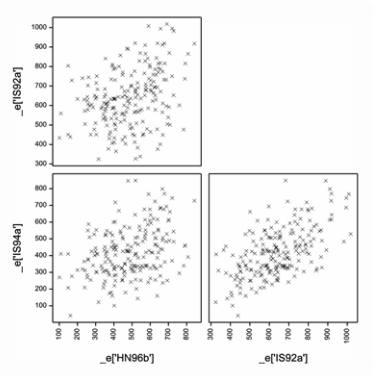
- Change of variation across environments
- Why?
  - different sets of active QTL/genes
  - Different intensity of action of QTL/genes
- Which environment show larger variation?




WAGENINGEN UR
For quality of life

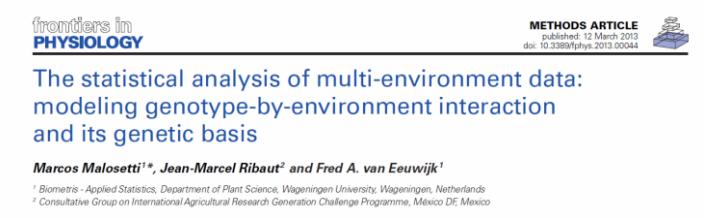
# Crude indications of GxE: 2. Lack of correlation

- Correlation between environments reflects GxE:
  - Positive high = no GxE
  - Positive but low = GxE
  - Negative = very strong GxE
- GxE → $r_g < 1$ → cannot precisely predict yield of a genotype in another environment




**Note:** In plants we can repeat genotypes (*e.g.,* clones or varieties), so we can directly "see" the genetic correlation.

## Example GxE Data: CIMMYT drought stress in maize



frontiers in PHYSIOLOGY

METHODS ARTICLE
published: 12 March 2013
doi: 10.3389/fphys.2013.00044

The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis

Marcos Malosetti[1]*, Jean-Marcel Ribaut[2] and Fred A. van Eeuwijk[1]

[1] Biometris - Applied Statistics, Department of Plant Science, Wageningen University, Wageningen, Netherlands
[2] Consultative Group on International Agricultural Research Generation Challenge Programme, México DF, Mexico

- Response variable: grain yield (ton/ha)
- 211 genotypes ($F_{2:3}$ lines).
- Eight environments: intermediate and severe drought stress (IS, SS), low and high nitrogen (LN, HN), no stress.
  - 1992, 1994, 1996
  - 2 locations (TI, PR)
  - Winter and summer seasons

WAGENINGEN UR
For quality of life

## Summary statistics related with GxE

- Differences in mean?
  - $\bar{x}_{IS94a} = 2.80 > \bar{x}_{LN94a} = 1.225$
- Differences in variation?
  - $sd_{IS94a} = 0.986 > sd_{LN94a} = 0.440\ ton/ha$
- The differences in variance hints at GxE.

Summary statistics for yield: E IS94a

| | |
|---|---|
| Number of observations = | 211 |
| Number of missing values = | 0 |
| Mean = | 2.80 |
| Median = | 2.743 |
| Minimum = | 0.278 |
| Maximum = | 5.658 |
| Lower quartile = | 2.125 |
| Upper quartile = | 3.406 |
| Standard deviation = | 0.986 |
| Variance = | 0.972 |

Summary statistics for yield: E LN96a

| | |
|---|---|
| Number of observations = | 211 |
| Number of missing values = | 0 |
| Mean = | 1.225 |
| Median = | 1.16 |
| Minimum = | 0.28 |
| Maximum = | 3.133 |
| Lower quartile = | 0.942 |
| Upper quartile = | 1.445 |
| Standard deviation = | 0.440 |
| Variance = | 0.194 |

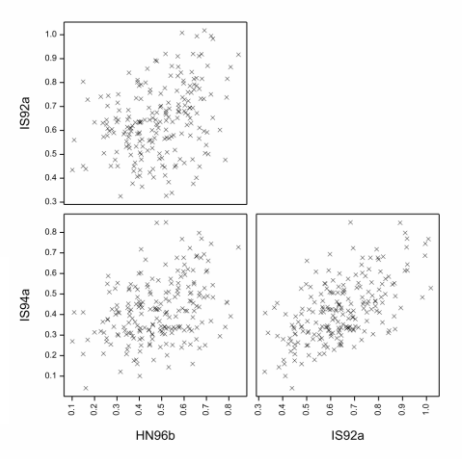
WAGENINGEN UR
For quality of life

# Co-variation between environments

- Correlation reveals GxE
  - $r = \frac{cov(x_1,x_2)}{\sqrt{Var(x_1)Var(x_2)}}$

Correlations

| | | | | |
|---|---|---|---|---|
| _e['HN96b'] | 1 | - | | |
| _e['IS92a'] | 2 | 0.3303 | - | |
| _e['IS94a'] | 3 | 0.3481 | 0.5877 | - |
| | | 1 | 2 | 3 |




WAGENINGEN UR
For quality of life

# 2. A taste of modelling GxE in plants

WAGENINGEN UR
For quality of life

## Statistical models for GxE data

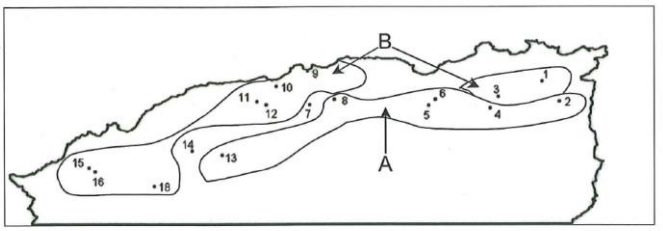
- GxE can be addressed by a combination of:
  - Linear models
  - Bi-linear models
  - Mixed models
- Linear-bilinear models with fixed effects:
  - Useful for exploratory analyses, but limited by
    - Model assumptions (eg: homoscedasticity)
    - Large number of parameters
- Mixed models: more natural to analyse MET data because
  - Heterogeneity of variances and co-variances in the data (between environments = GxE).
  - Model heterogeneity of variation within individual trials (including global and local spatial trends).

## One vs two-stage analyses

- One-stage analysis:
  - Simultaneous modelling of the within and between trial variation.
  - Uses all the information (plot data).
  - Complicated (to account individual trials specifics).
- Two-stage analysis:
  - Stage 1: analysis of individual trials; design issues in a particular trial.
  - Stage 2: collate adjusted means obtained from stage 1, to run a joint GxE analysis. Now the focus is on GxE.
  - Pragmatic approach with little loss of information. (weighted analysis if needed).

WAGENINGEN UR
For quality of life

# Example durum wheat Algeria

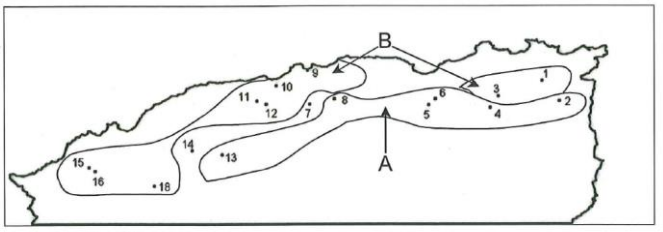



*Source:* Annicchiarico, 2002c.

- 24 genotypes in 22 environments
  - 11 sites
  - 2 years
- Experiments designed as RCBD in each site.
  - RCBD = "randomized complete block design"

WAGENINGEN UR
For quality of life

# One-stage analyses




*Source:* Annicchiarico, 2002c.

WAGENINGEN UR
For quality of life

## One-stage analysis: basic ANOVA model

$$\underline{y}_{ijk} = \mu + G_i + E_j + b_{k(j)} + GE_{ij} + \underline{e}_{ijk} \qquad \underline{e}_{ijk} \sim N(0, \sigma^2)$$

- Grain yield =
  - Intercept +
  - Genotype i ($G_i$) +
  - Environment j ($E_j$) +
  - Block within environment ($b_{k(j)}$)
  - Genotype x Environment interaction ($GE_{ij}$) +
  - error ($\underline{e}_{ijk}$)

- Linear model (fixed effects model):
  - One parameter per genotype, environment, and combination of genotype and environment. (Hence: many parameters)
  - Requires balanced data (at least for easy interpretation).
  - Assumes constant residual variance.

## Classical ANOVA results

Analysis of variance

Variate: yield

| Source of variation | d.f. | s.s. | m.s. | v.r. | F pr. |
|---|---|---|---|---|---|
| Block.Environment stratum | | | | | |
| Environment | 21 | 935.94627 | 44.56887 | 25.00 | <.001 |
| Residual | 66 | 117.68513 | 1.78311 | 21.49 | |
| Block.Environment.*Units* stratum | | | | | |
| Genotype | 23 | 80.84161 | 3.51485 | 42.36 | <.001 |
| Environment.Genotype | 483 | 151.66288 | 0.31400 | 3.78 | <.001 |
| Residual | 1518 | 125.94926 | 0.08297 | | |
| Total | 2111 | 1412.08514 | | | |

- Environment effects often largest.
- Difficult to compare variation genotype main effects versus genotype by environment interaction.
- When genotypes are just a sample from a larger population, classical ANOVA is not the most useful
  - Full parameterization does not teach us much...
  - → switch to random genotypic effects, mixed model.

## One-stage analysis: A basic mixed model

$$\underline{y}_{ijk} = \mu + \underline{G}_i + E_j + b_{k(j)} + \underline{GE}_{ij} + \underline{e}_{ijk}$$

$$\underline{G}_{ijk} \sim N(0, \sigma_G^2)$$
$$\underline{GE}_{ijk} \sim N(0, \sigma_{GE}^2)$$
$$\underline{\epsilon}_{ijk} \sim N(0, \sigma^2)$$

- Grain yield =
  - Intercept +
  - Effect of genotype i (Gi) +
  - Effect of environment j ($E_j$) +
  - Block within environment ($b_{k(j)}$)
  - Effect of genotype x environment interaction (GEij) +
  - error ($\underline{e}_{ijk}$)
- Linear mixed model:
  - One parameter for G and one for GxE → variance components.
  - No problem with unbalanced data.
  - Allows to estimate different variance components (also for residuals).
- This is still a very simple (unrealistic) model → compound symmetry model

## Estimates of variance components

- $\hat{\sigma}_G^2 = 0.036; \quad \hat{\sigma}_{GE}^2 = 0.058$
- $\hat{\sigma}^2 = 0.083$
- $\hat{\sigma}_{GE}^2$ larger than $\hat{\sigma}_G^2$
- $r = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2 + \sigma^2} = 0.205$ ⇦ repeatability
  - ratio $\hat{\sigma}_G^2 / \hat{\sigma}_{GE}^2$=0.63 (next slide)

**Observations:**

- Var(GxE) almost double var(G)
  - A lot of the genetic variation is specific to environments
- GxE is important.
  - Relative high with respect to main effect
  - Low correlation between environments (r) → high GxE

WAGENINGEN UR
For quality of life

## Comparison with other examples (from literature)

| Crop | Region | Vg | Vgxe | Ve | Vg/Vge |
|---|---|---|---|---|---|
| Spring Barley | Canada | 62 | 110 | 174 | 0.56 |
| Spring Oat | Canada | 122 | 132 | 178 | 0.92 |
| Wheat | Australia | 23 | 70 | 87 | 0.33 |
| Winter wheat | UK | 99 | 142 | 128 | 0.70 |
| Potatoes | UK | 9780 | 20570 | 18790 | 0.48 |
| Lowland rice | Thailand | 198 | 299 | 178 | 0.66 |
| Lowland rice | Thailand | 60 | 311 | 440 | 0.19 |
| durum wheat | Algeria | 0.0364 | 0.0578 | 0.0830 | 0.63 |

- Often GxE > G (ratio lower than 0.5).
- Relatively high GxE (Spring Oat example of relatively low GxE)

WAGENINGEN UR
For quality of life

## Towards structure: Partitioning of GxE

$$\underline{y}_{ijk} = \mu + \underline{G}_i + \underline{E}_j + \underline{GE}_{ij} + \underline{e}_{ijk}$$

$$\underline{y}_{ijk} = \mu + \underline{G}_i + L_m + Y_n + LY_{mn} + \underline{GL}_{im} + \underline{GY}_{in} + \underline{GLY}_{imn} + \underline{e}_{ijk}$$

- Environments can be partitioned into components:
  - Locations (e.g.: geography, soils, topology, etc)
  - Years (e.g.: weather conditions)
  - Locations x Years (combination of both)
- Useful to split GxE into components: GxL, GxY, GxLxY
  - To interpret GxE (what might have caused it?)
  - To design network of evaluation trials (define number of years/ locations/ replicates)

## Partitioning of GxE & predictability

$$\underline{y}_{ijk} = \mu + \underline{G}_i + L_m + Y_n + LY_{mn} + \underline{GL}_{im} + \underline{GY}_{in} + \underline{GLY}_{imn} + \underline{e}_{ijk}$$

- Understanding causes of GxE:
  - GxL: differential genotypic response to conditions that **are** particular to the location
  - GxY: differential genotypic response to conditions that **were** particular to the year/season
  - GxLxY: differential genotype response to conditions that **were** particular to the location and year/season combination.
- Which causes are more likely to be repeatable?
- Structuring the TPE → obtain predictability

## Partition of GxE: GxL, GxY, and GxLxY

$$\underline{y}_{ijk} = \mu + \underline{G}_i + \underline{E}_j + \underline{GE}_{ij} + \underline{e}_{ijk}$$

$$\underline{y}_{ijk} = \mu + \underline{G}_i + L_m + Y_n + LY_{mn} + \underline{GL}_{im} + \underline{GY}_{in} + \underline{GLY}_{imn} + \underline{e}_{ijk}$$

- $\underline{GE}_{ij}$ partitioned into: GL, GY, GLY
  - $\sigma^2_{GE} = \sigma^2_{GL} + \sigma^2_{GY} + \sigma^2_{GLY}$
- Quantify the predictable component of GxE (i.e. $\sigma^2_{GL}$)

WAGENINGEN UR
For quality of life

# Resulting variance components

**Estimated variance components**

| Random term | component | s.e. |
|---|---|---|
| Genotype | 0.03516 | 0.01186 |
| Genotype.Location | 0.00481 | 0.00513 |
| Genotype.Year | 0.00207 | 0.00264 |
| Genotype.Location.Year | 0.05209 | 0.00683 |

**Residual variance model**

| Term | Model(order) | Parameter | Estimate | s.e. |
|---|---|---|---|---|
| Residual | Identity | Sigma2 | 0.0830 | 0.00301 |

- In the previous analysis: $\hat{\sigma}^2_{GE} = 0.058$
- Now: $\hat{\sigma}^2_{GL} = 0.005$, $\hat{\sigma}^2_{GY} = 0.002$, and $\hat{\sigma}^2_{GLY} = 0.052$
- GxLxY the most important (difficult, non-repeatable GxE)
  - Very limited predictability of the GxE here.
- Note that: 0.005+0.002+0.052 ≅ 0.058

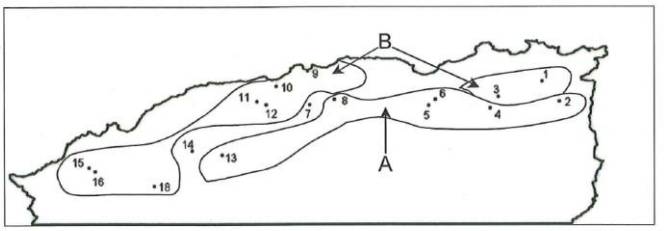
# Importance of GxL, GxY, and GxLxY

| Crop | Region | Vg | Vgxl | Vgxy | Vgxlxy | Ve | Vgxl/Vgxe |
|---|---|---|---|---|---|---|---|
| Spring Barley | Canada | 62 | 29 | 18 | 63 | 174 | 0.26 |
| Spring Oat | Canada | 122 | 58 | 21 | 53 | 178 | 0.44 |
| Wheat | Australia | 23 | 8 | 9 | 53 | 87 | 0.11 |
| Winter wheat | UK | 99 | 7 | 22 | 113 | 128 | 0.05 |
| Potatoes | UK | 9780 | 2980 | 2630 | 14960 | 18790 | 0.14 |
| Lowland rice | Thailand | 198 | 82 | 18 | 199 | 178 | 0.27 |
| Lowland rice | Thailand | 60 | 3 | 49 | 259 | 440 | 0.01 |
| durum wheat | Algeria | 0.0364 | 0.0048 | 0.0021 | 0.0521 | 0.083 | 0.08 |

- Often limited predictability of GxE
  - Var(GxL)/var(GXE) << 1
  - Spring oat an example of relatively high "repeatable" GxE.

WAGENINGEN UR
For quality of life

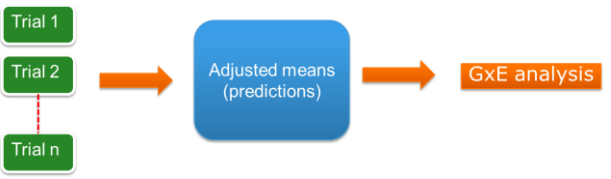
# Two-stage analysis




*Source:* Annicchiarico, 2002c.

WAGENINGEN UR
For quality of life

# Two-stage analysis




- First stage: analysis per trial
  - Quality control (assumptions/outliers/etc).
  - Obtain adjusted means (and weights) per genotype.
  - Focus on the within trial modelling: incomplete blocks, row-column effects, local effects (spatial modelling), etc
- Second stage: use adjusted means in GxE analysis:
  - Model GxE in terms of genotypic and/or environmental specific parameters.

## Second stage: Modelling of GxE

$$\underline{y}_{ij} = \mu + G_i + E_j + \left(\underline{GE}_{ij} + \underline{e}_{ij}\right)$$

- The GxE modelling at the level of the GxE table of means:
- Note that $\underline{GE}_{ij}$ is part of the residual (lowest level of the data):
- The full factorial ANOVA model is not of main interest:
  - It does not teach us much about GxE.
  - It has limited prediction ability.
  - ...but it is the starting point in the GxE modelling exercise
- Main task is to model the GxE in terms of parameters that are either genotypic or environmental specific:
  - Exclusive dependence on genotype / environment = separability
  - Separability implies we can control / adjust / modify the response

## Snapshot of models for GxE (the mean)

$$\underline{y}_{ij} = \mu + G_i + E_j + \left(\underline{GE}_{ij} + \underline{e}_{ij}\right) \quad \text{Full factorial ANOVA}$$

$$\underline{y}_{ij} = \mu + G_i + E_j + \beta_i E_j + \underline{\epsilon}_{ij} \quad \text{Finlay Wilkinson}$$

$$\underline{y}_{ij} = \mu + G_i + E_j + \sum_{k \in K} \lambda_k u_{ik} v_{jk} + \underline{\epsilon}_{ij} \quad \text{AMMI}$$

$$\underline{y}_{ij} = \mu + E_j + \sum_{k \in K} \lambda_k u_{ik} v_{jk} + \underline{\epsilon}_{ij} \quad \text{GGE}$$

$$\underline{y}_{ij} = \mu + G_i + E_j + x_i \alpha_j + \beta_i Z_j + \underline{\epsilon}_{ij} \quad \text{Factorial regression}$$

- Examples of different models (to be discussed later)
- Note that parameters are either "red" (genotypic) or "blue" (environmental) → separability

WAGENINGEN UR
For quality of life

# Summary

- Simple summary statistics: indications of GxE.

- Modelling GxE:
  - ANOVA and (particularly) mixed models → useful starting points in GxE-analysis.
  - More elaborate modelling is needed →
  - find good models with "genotypic" and/or "environmental" specific parameters → separability.

- To be continued...

WAGENINGEN UR
For quality of life