strube
Breeding progress together

UNIVERSITY OF HOHENHEIM

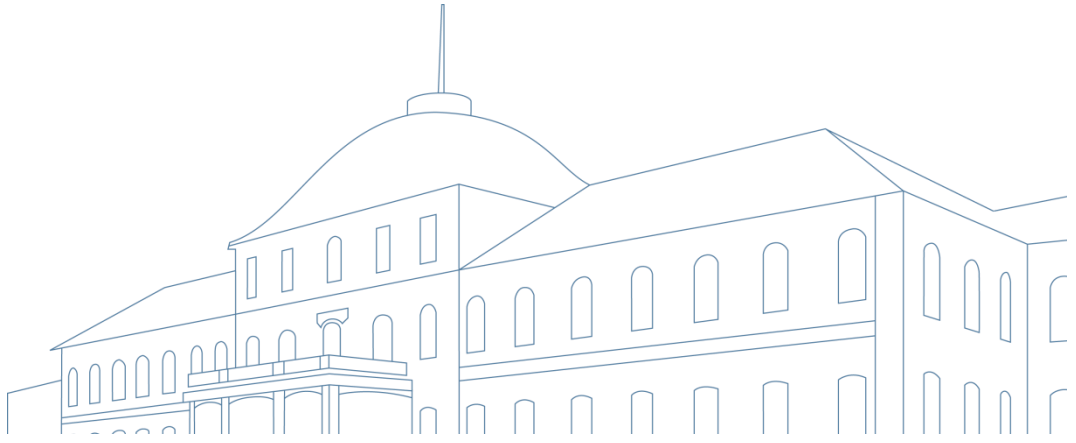
Environmental information for yield prediction in sugar beet breeding: the best-case scenario

Maksym Hrachov[1], Elena Orsini[2], Bettina Müller[2], Jinquan Li[2], Hans-Peter Piepho[1]

[1] Biostatistics unit, University of Hohenheim

[2] Strube D&S GmbH / RAGT

Correspondence:
maksym.hrachov@uni-hohenheim.de

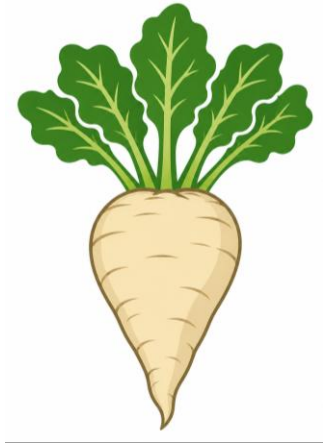


Content

1. Project aims
2. Challenges of multi-environmental trials
3. Statistical models
4. Results & outlook

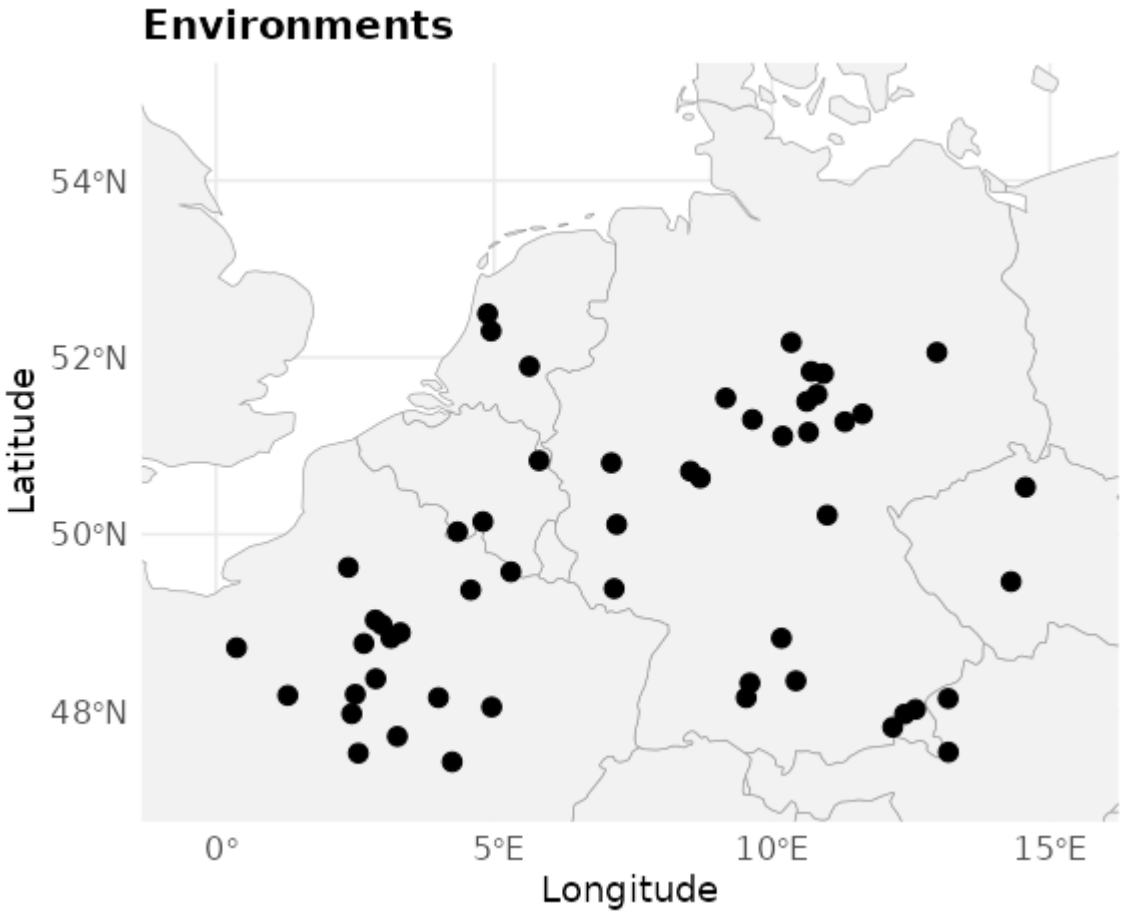
Project aims

This is a cooperation between Strube D&S GmbH and the University of Hohenheim



- use data from the sugar beet variety development (plant breeding)
- adapt or design models that improve over the routine model
  - that includes raw field and genomic data analysis, environmental variables definition, etc ...

Challenges of multi-environmental trials



Multi-environmental trial (MET):

- 2 years
- 53 environments (location by year)
- ~3500 genotypes (test subjects)
- 6 check varieties (control)
- 100...3000 entries per environment
- connectivity is achieved via replicated check varieties (15% of total entries)

Disclaimer: this map does not reflect the true positions, but only simulates how the distribution of environments may look like

The best-case scenario

- Tested genotypes have low connectivity
- Check varieties are highly replicated in each environment

- Focus on check varieties only to establish the best-case scenario
  - ➢ This gives reduced data size, thus an opportunity to try many more approaches
- Quantify maximum achievable gain
- See how much of it can be recovered using complete data

Challenges of multi-environmental trials

- ☑ Logistical complexity (dimensions, size)
- ☑ Connectivity of test subjects
- ❑ Individuality (field structure and condition)

Challenges of multi-environmental trials

Procedure:

First stage analysis
(individual field model)

↓

Adjusted means and their variances

↓

Second stage model

Benefits:

Per-environment analysis

+

Reduced data size

+

Faster model convergence

First-stage analysis

$$\begin{Bmatrix} g_i\,, \\ g_i + r(d)\,, \\ g_i + k(c)\,, \\ g_i + r(d) + k(c) \end{Bmatrix} + \begin{Bmatrix} 0\,, \\ b_s\,, \\ b_s + r'_d\,, \\ b_s + k'_c\,, \\ b_s + r'_d + k'_c\,, \\ b_s + r_{\text{spline}}(d)\,, \\ b_s + k_{\text{spline}}(c)\,, \\ b_s + r_{\text{spline}}(d) + k_{\text{spline}}(c) \end{Bmatrix} + \begin{Bmatrix} e \sim N(0, \sigma_e^2)\,, \\ e \sim N(0, \sigma_e^2(\text{AR1}_{\text{row}} \otimes \text{AR1}_{\text{col}})) \end{Bmatrix}$$

Fixed Random Residual

First-stage analysis

$$\begin{Bmatrix} g_i \\ g_i + r(d) \\ g_i + k(c) \\ g_i + r(d) + k(c) \end{Bmatrix} + \begin{Bmatrix} 0 \\ b_s \\ b_s + r'_d \\ b_s + k'_c \\ b_s + r'_d + k'_c \\ b_s + r_{\text{spline}}(d) \\ b_s + k_{\text{spline}}(c) \\ b_s + r_{\text{spline}}(d) + k_{\text{spline}}(c) \end{Bmatrix} + \begin{Bmatrix} e \sim N(0, \sigma_e^2) \\ e \sim N(0, \sigma_e^2(\text{AR1}_{\text{row}} \otimes \text{AR1}_{\text{col}})) \end{Bmatrix}$$

Genotype

Linear row/column effects

Design block

Factor row/column or splines for row/column

Independent error or autoregressive for row/column

* If “Field Block” is present, most model terms are nested within it;

First-stage analysis

$$\begin{Bmatrix} g_i\,, \\ g_i + r(d)\,, \\ g_i + k(c)\,, \\ g_i + r(d) + k(c) \end{Bmatrix} + \begin{Bmatrix} 0\,, \\ b_s\,, \\ b_s + r'_d\,, \\ b_s + k'_c\,, \\ b_s + r'_d + k'_c\,, \\ b_s + r_{\text{spline}}(d)\,, \\ b_s + k_{\text{spline}}(c)\,, \\ b_s + r_{\text{spline}}(d) + k_{\text{spline}}(c) \end{Bmatrix} + \begin{Bmatrix} e \sim N(0, \sigma_e^2)\,, \\ e \sim N(0, \sigma_e^2(\text{AR1}_{\text{row}} \otimes \text{AR1}_{\text{col}})) \end{Bmatrix}$$

50 models

Average reduction of 57 AIC* units (best model vs. baseline)

*Akaike information criterion: evaluates log-likelihood fit of a model, and adds penalty based on the number of model parameters.

Available data

- Check variety information (5-6 per environment)

Resulting dataset: 53 environments, 291 adjusted means.

- agERA5 environmental data (6 EC)

For the growing season: rainfall, solar radiation, temperature, growing degree days, relative humidity, evapotranspiration.

- Breeding zones (7 clusters based on breeder's knowledge)

Climate and disease-based clustering.

Models

Baseline: CSY ~ g + loc + year + (loc*year) + (g*loc) + (g*year) + (g*loc*year)

CSY ~ g + e + (g*e)

All these effects are random!

CSY – commercial sugar yield
g – genotype
loc – location
year – year
e – environment
(loc*year) – interaction notation

Models

Baseline: CSY ~ g + e + (g*e)

Model 1*: CSY ~ e + (g*e) + zone + (g*zone)

Model 2*: CSY ~ g + e + (g*e) + EC + (g*EC)

Model 3*: CSY ~ e + (g*e) + zone + EC + (g*zone) + (g*EC)

Modelling these two components jointly is a big challenge!

g – genotype

e – environment

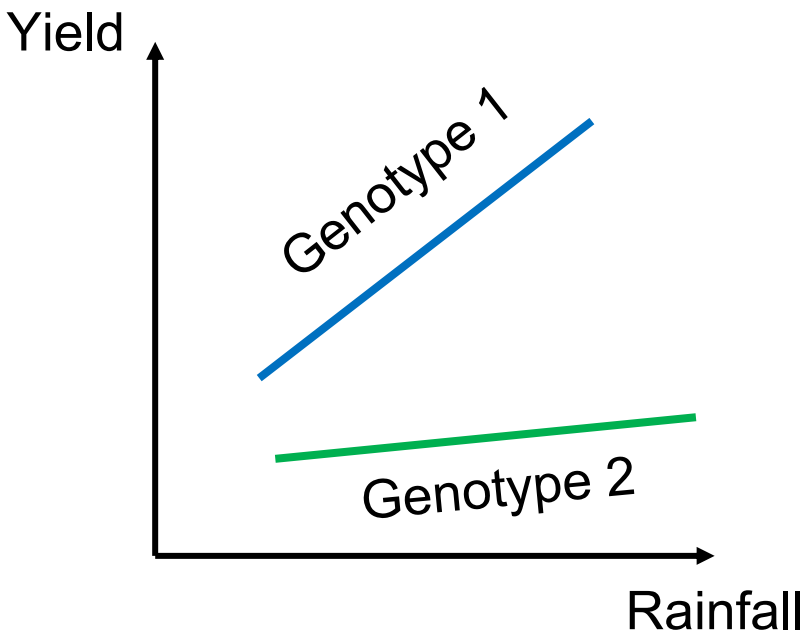
(g*e) – specifies an interaction of g and e

EC – environmental covariates

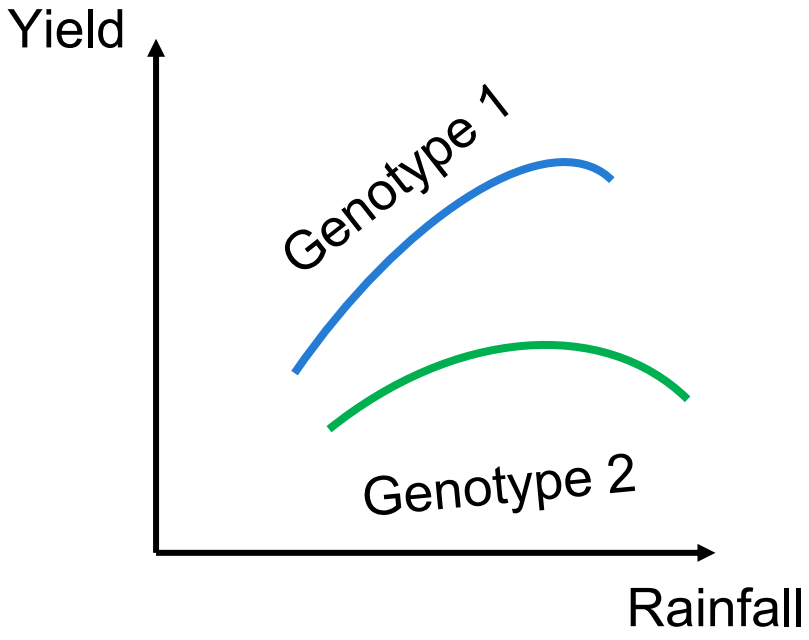
* - models 1-3 involved either unstructured or reduced rank variance-covariance spanning g*zones or g+g*EC

EC specification: linear vs. quadratic

Linear

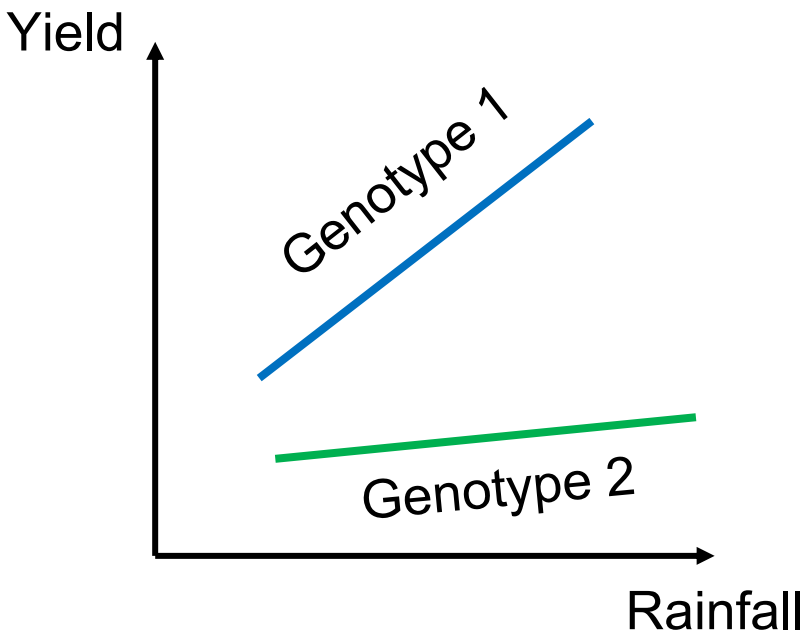


Quadratic

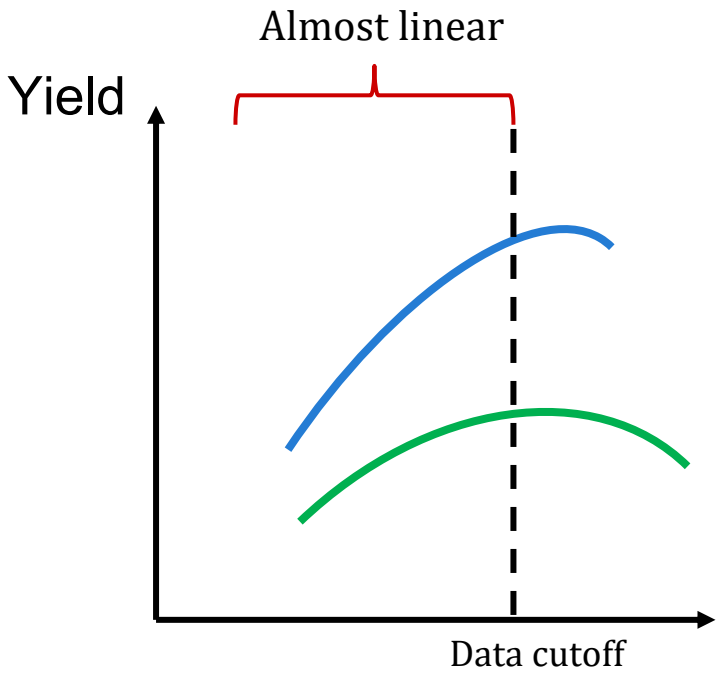


Polynomial effects: we don't always have the data

Linear



Quadratic



Model evaluation: cross-validation

Baseline: CSY ~ g + e + (g*e)
Model 1*: CSY ~ e + (g*e) + zones + (g*zones)
Model 2*: CSY ~ g + e + (g*e) + EC + (g*EC)
Model 3*: CSY ~ e + (g*e) + zones + EC + (g*zones) + (g*EC)

g – genotype
e – environment
(g*e) – specifies an interaction of g and e
EC – environmental covariates
Note: models 1-3 involved either unstructured or reduced rank variance-covariance spanning g*zones or g+g*EC

$$MSEPD_j = \frac{2\sum_{i=1}^{I}\left(f_{ij} - \bar{f}_{\bullet j}\right)^2}{(I-1)}$$

Shows the error of genotype differences – useful for breeding

$$MSPE_j = \frac{\sum_{i=1}^{I}\left(f_{ij}\right)^2}{I}$$

Shows the mean error – useful for prediction on crop level

Given $f_{ij} = y_{ij} - \hat{y}_{ij}$

Where y_{ij} is the adjusted mean,
and $\hat{y}_{ij}$ is the predicted value

Conclusions so far

1. Inclusion of zones alone brings a large breeding-relevant improvement (disease influence?)

2. EC need some selection and careful choice before visible MSEPD improvement is achieved.

3. EC bring large MSPE improvement, more relevant for crop-level prediction than breeding.

4. A combination of zone and EC information has the highest impact on prediction performance.

Thank you for your attention!

Let’s discuss

Sources of external images

- University of Hohenheim logo, all slides: https://www.uni-hohenheim.de/en
- Strube Gmbh logo, all slides: https://www.strube.net
- Sugar beet field: ChatGPT image generation
- Cooperation industry-academia: ChatGPT image generation
- Map of environments – simulation of 53 locations in several EU countries, does not reflect the true positions.