Assessing phenotypic stability in plant breeding trials

Hans-Peter Piepho

Biostatistics Unit

University of Hohenheim

Germany

Content

1. Introduction
2. A few stability measures
3. Statistical models
4. Sample size
5. Additive Main Effects Multiplicative Interaction (AMMI)
6. Probability methods
7. Three-factorial analysis of series of trials
8. Making use of environmental covariate information
9. Time trend for stability
10. Concluding remarks

1. Introduction

- Plant breeding and cultivar trials
- Trial series over several environments (sites, years)
- Individual trials replicated

  (randomized complete block designs, lattice designs, α-Designs)
- Often analyses in two stages:

  (i) Genotype means per environment

  (ii) Analysis og genotype-environment means by factorial model

Objective:

- Identify genotypes with best mean yield
- Identify genotypes with best "yield stability"

2. A few stability measures

Environmental variance

$$S_i^2 = \frac{\sum_{j=1}^{J} \left(y_{ij} - \bar{y}_{i\bullet} \right)^2}{J-1} \quad ,$$

where

y_{ij} = yield of *i*-th genotype in *j*-th environment

(Römer, 1917)

Ecovalence

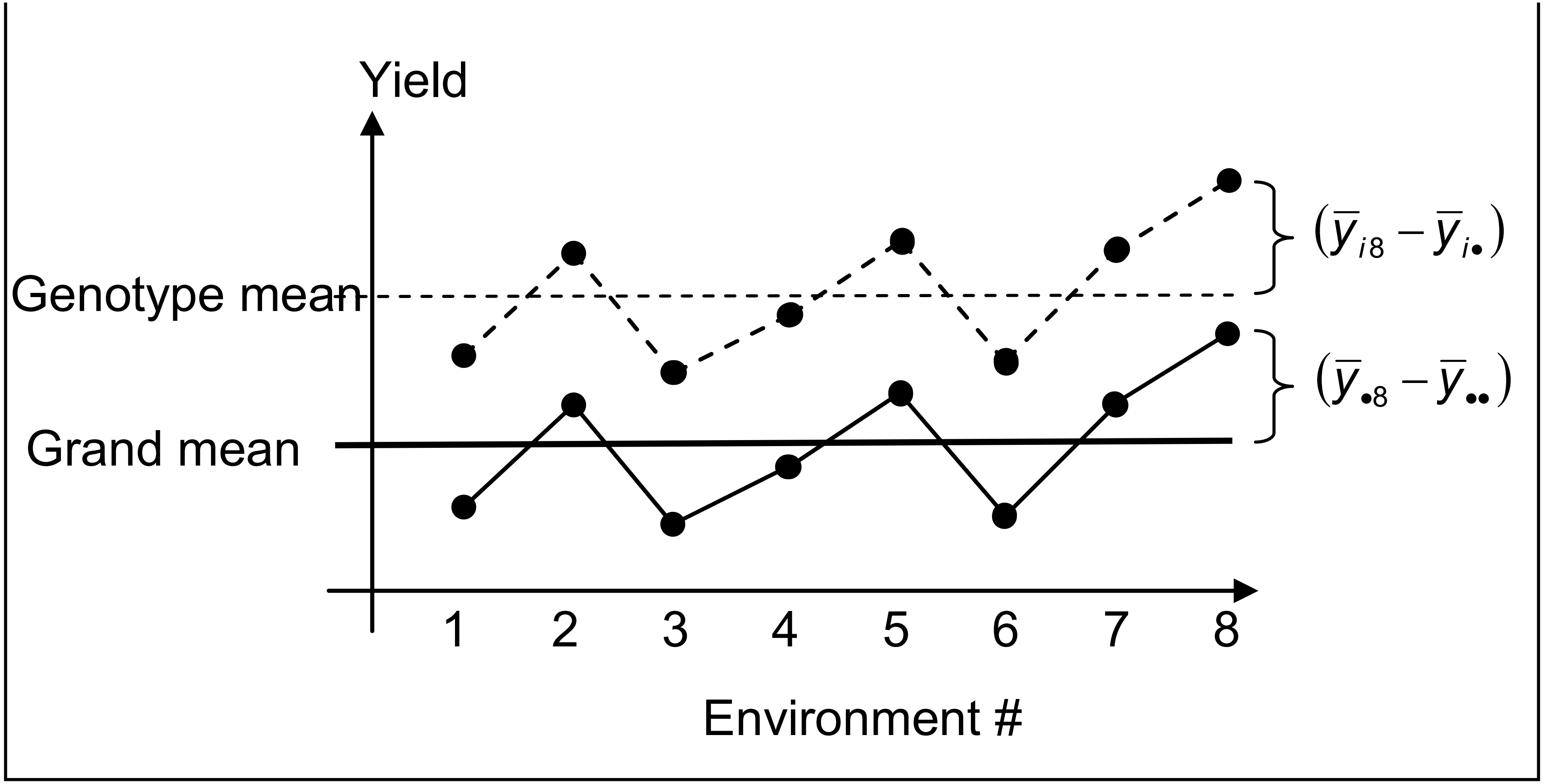


Fig. 1: Schematic representation of a stable genotype with yield following the environmental mean.

Deviation of genotype from its mean across environments: $\left(y_{ij} - \bar{y}_{i\bullet}\right)$

Deviation of environmental mean from grand mean: $\left(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}\right)$

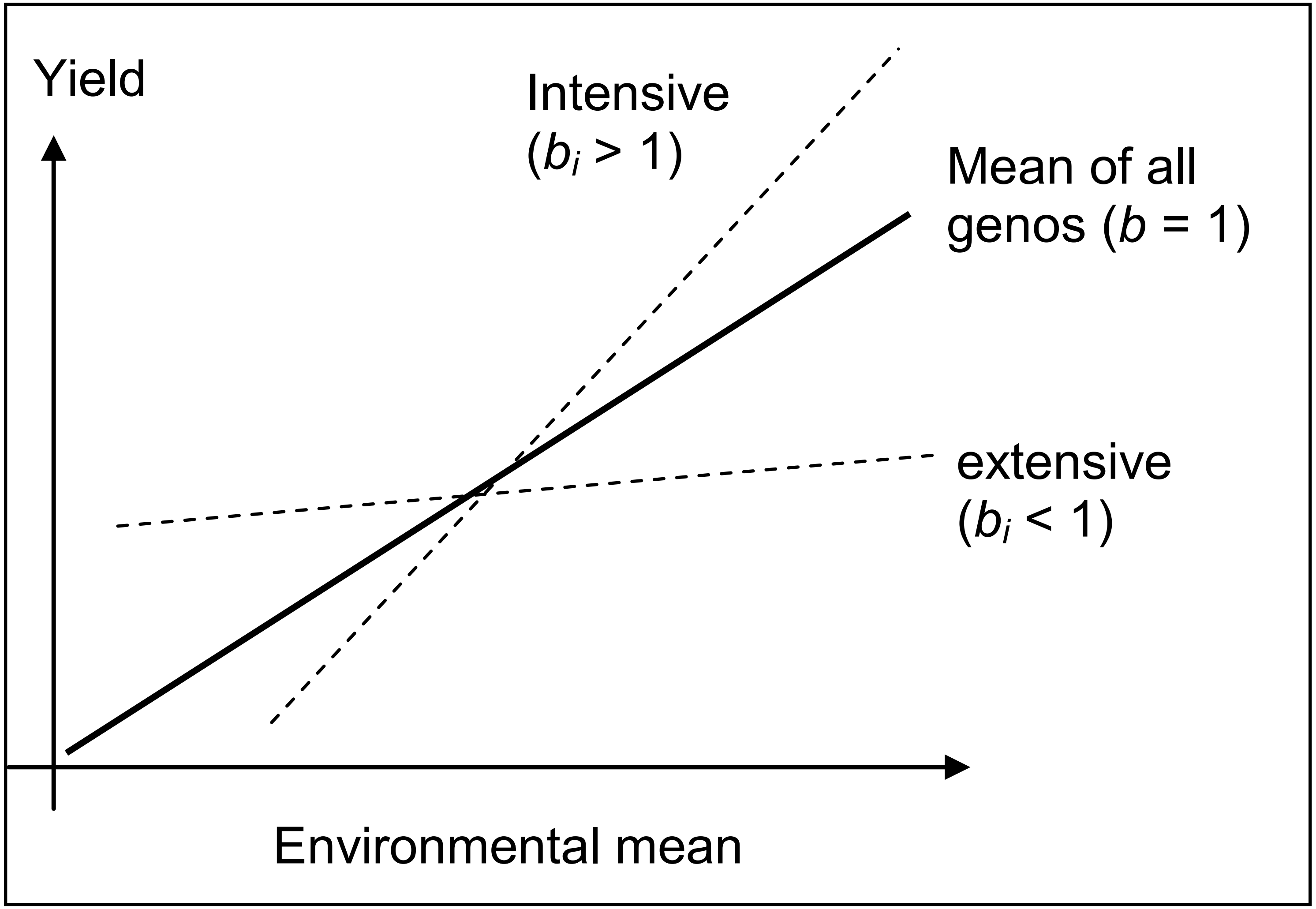
[= Mean of $\left(y_{ij} - \bar{y}_{i\bullet}\right)$ across genotypes!]

Difference: $r_{ij} = \left(y_{ij} - \bar{y}_{i\bullet}\right) - \left(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}\right) = \left(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}\right)$

$$W_i = \sum_{j=1}^{J} \left(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet}\right)^2$$

(Tadeusz Caliński, 1960; Günter Wricke, 1962)

Regression approach



Regression coefficient:

b_i = 1 (Genotype shows average response) $\Rightarrow$ stable <u>or</u>

b_i = 0 (Genotype shows no response at all) $\Rightarrow$ stable

Deviations from regression:

$s^2_{d(i)}$ = Variance of deviations from regression = 0 $\Rightarrow$ stable

(Finlay and Wilkinson, 1963; Eberhart and Russell, 1966)

3. Statistical models

2-factorial ANOVA model

$y_{ij} = \mu + g_i + e_j + (ge)_{ij}$,

where

g_i = main effect of *i*-th genotype

e_j = main effect of *j*-th environment

$(ge)_{ij}$ = Interaction of *i*-th genotype with *j*-th environment

Ecovalence ⇔ Stability variance = $\sigma^2_{ge(i)} = var\left[(ge)_{ij}\right]$

(Wricke, 1962; Shukla, 1972)

Regression approach

$y_{ij} = \mu + g_i + \beta_i e_j + d_{ij}$,

where

β_i = regression coefficient of *i*-th genotype

d_{ij} = deviation of *i*-th genotype in *j*-th environment

Variance of deviations d_{ij} as stability measure:

$$\sigma^2_{d(i)} = var\left(d_{ij}\right)$$

Stability measures b_i and $s^2_{d(i)}$ are estimators of parameters β_i and $\sigma^2_{d(i)}$

Environmental variance

$y_{ij} = \mu + g_i + f_{ij}$.

$\sigma_{ii} = \mathrm{var}(f_{ij})$

The environmental variance S_i^2 is an estimator of σ_{ii}

Structure of residual effect f_{ij}

Environmental variance:

none!

Stability variance:

$f_{ij} = e_j + (ge)_{ij}$

Regression approach:

$f_{ij} = \beta_i e_j + d_{ij}$

3.1 Random environments, fixed genotypes

Environments:

- Target population of environments (TPE)
- Need representative sample of environments from TPE

Without random sampling of environments stability analysis is of limited value.

⇒ Environments random

Genotypes:

- Objective: Estimation of mean performance of every genotype
- Caracterization of population of genotypes of no interest

$\Rightarrow$ Genotypes fixed

But:

- In some circumstances better to take genotypes as random

  $\Rightarrow$ Use of pedigree or kinship information in breeding program

- Best Linear Unbiased Prediction (BLUE) for fixed effects tends to be less precise than Best Linear Unbiased Prediction (BLUP) for random effects

3.2 Variance-covariance structures for genotype-environment data

2-factorial ANOVA:

$y_{ij} = \mu + g_i + e_j + (ge)_{ij}$

where

$$e_j \sim N\left(0, \sigma_e^2\right)$$

and

$$(ge)_{ij} \sim N\left(0, \sigma_{ge}^2\right)$$

Variance of an observation: $var\left(y_{ij}\right) = \sigma_e^2 + \sigma_{ge}^2$

Covariance among 2 genotypes: $cov\left(y_{ij}, y_{i'j}\right) = \sigma_e^2$

Stability variance:

$$(ge)_{ij} \sim N\left(0, \sigma^2_{ge(i)}\right)$$

Variance of an observation: $var\left(y_{ij}\right) = \sigma^2_e + \sigma^2_{ge(i)}$

Covariance among 2 genotypes: $cov\left(y_{ij}, y_{i'j}\right) = \sigma^2_e$

Regression approach:

$y_{ij} = \mu + g_i + \beta_i e_j + d_{ij}$

with

$e_j \sim N\left(0, \sigma_e^2\right)$ and

$d_{ij} \sim N\left(0, \sigma_{d(i)}^2\right)$

Variance of an observation: $var\left(y_{ij}\right) = \beta_i^2 \sigma_e^2 + \sigma_{d(i)}^2$

Covariance among 2 genotypes: $cov\left(y_{ij}, y_{i'j}\right) = \beta_i \beta_{i'} \sigma_e^2$

Environmental variance:

$y_{ij} = \mu + g_i + f_{ij}$

Variance of an observation: $var\left(y_{ij}\right) = var\left(f_{ij}\right) = \sigma_{ii}$

Covariance among 2 genotypes: $cov\left(y_{ij}, y_{i'j}\right) = cov\left(f_{ij}, f_{i'j}\right) = \sigma_{ii'}$

Tab. 1: Variances and covariances for different models (§: *I* = no. of genotypes).

Model	$var(y_{ij})$	$cov(y_{ij}, y_{i'j})$	no. of parameters§
2-factorial ANOVA	$\sigma_e^2 + \sigma_{ge}^2$	σ_e^2	2
Stability variance	$\sigma_e^2 + \sigma_{ge(i)}^2$	σ_e^2	*I* + 1
Regression approach	$\beta_i^2\sigma_e^2 + \sigma_{d(i)}^2$	$\beta_i\beta_{i'}\sigma_e^2$	2*I*
Environmental variance	σ_{ii}	$\sigma_{ii'}$	*I*(*I*+1)/2

3.3 Estimation of stability and model selection

Tab. 2: Unbalanced dataset from Piepho (1999) (Dataset 1).

	Environment																									
Genotype	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	.	.	.	.	.	.	.	.	.	63	72	69	55	52	67	63	55	66	57	69	73	68	64	67	70	70
2	.	.	.	.	.	.	.	.	.	64	82	76	53	59	62	74	54	66	54	64	69	66	79	66	66	65
3	.	.	.	.	75	72	55	52	79	72	83	78	40	54	65	73	.	.	.	.	.	.	.	.	.	.
4	66	59	67	55	85	77	51	49	80	65	88	86	63	53	68	63	52	.	.	.	.	.	.	.	.	60
5	49	40	57	53	74	71	46	44	62	51	72	70	60	51	65	59	51	66	.	.	.	.	.	.	.	.

Estimating the models

⇒ Restricted Maximum Likelihood (REML) method

(SAS, GENSTAT, ASREML, R)

Model selection

Akaike Information Criterion (AIC):

$$AIC = -2LL_{REML} + 2p\ ,$$

where

LL_{REML} = REML log-likelihood

p = number of variance parameters

The smaller the value of AIC the better the model fit

Tab. 3: Fit of different models for Dataset 1.

Model	*p*[a]	AIC[b]
2-factorial ANOVA	2	–273.37
Stability variance	6	–275.26
Regression approach	10	–275.27
Environmental variance	15	–269.94

(a) *p* = number of variance parameters

(b) AIC = Akaike Information Criterion

Tab. 4: Parameter estimates (REML) with standard errors (s.e.) for variance parameters with different stability models (Datenset 1).

Genotype	Stability model							
	Stability variance		Regression approach				Environ. variance	
	$\sigma^2_{ge(i)}$	(s.e.)	β_i	(s.e.)	$\sigma^2_{d(i)}$	(s.e.)	σ_{ii}	(s.e.)
1	12.17	(11.94)	5.22	(1.51)	16.92	(8.13)	42.36	(13.45)
2	23.77	(15.09)	7.78	(1.92)	17.31	(15.60)	78.90	(25.56)
3	54.36	(27.02)	9.61	(2.37)	43.81	(29.47)	159.04	(57.42)
4	39.25	(18.00)	10.79	(2.05)	19.55	(19.34)	136.15	(40.63)
5	22.95	(12.81)	8.03	(1.81)	29.73	(15.25)	99.66	(32.19)

Convergence problems with REML

⇒ too few environments!

⇒ more genotype than environments

⇒ model too complex

⇒ limit attention to stability variance and regression approach

⇒ model environmental main effect as fixed (no recovery of inter-environment information)

4. Sample size

S_i^2 is an estimator of the true environmental variance σ_{ii}

Coefficient of variation:

$$CV\left(S_i^2\right) = \frac{\sqrt{var\left(S_i^2\right)}}{E\left(S_i^2\right)} = \sqrt{2/(J-1)}$$

J = number of environments

J = 50 $\Rightarrow CV\left(S_i^2\right) = 20\%$

J = 200 $\Rightarrow CV\left(S_i^2\right) = 10\%$

5. Additive Main Effects Multiplicative Interaction (AMMI)

$y_{ij} = \mu + g_i + e_j + \lambda_{i1} w_{j1} + \lambda_{i2} w_{j2} + \ldots + d_{ij}$

where

λ_{ik} = coefficients (scores) for genotypes

w_{jk} = coefficients (scores) for environments , $w_{jk} \sim N(0, 1)$

⇒ extension of regression approach

⇒ factor environment random: very flexible variance-covariance structure

⇒ factor-analytic models: $\mathrm{var}\left(y_{1j}, \ldots, y_{Ij}\right) = \Lambda\Lambda^T + D, \Lambda = \left\{\lambda_{ik}\right\}, D = diag\left(\sigma_{d1}^2, \ldots, \sigma_{dI}^2\right)$

⇒ explorative analysis of genotype-environment interactions using biplots

Tab. 5: Dataset 2.

	Environment																
Genotype	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	2.70	2.32	2.35	1.86	4.76	5.13	2.37	3.18	3.60	3.99	2.51	4.71	2.46	2.98	4.06	2.55	4.10
2	2.77	2.56	2.65	2.03	4.77	4.24	2.31	3.27	3.33	3.86	3.25	4.10	2.97	2.91	4.25	2.35	3.95
3	3.13	3.72	3.47	2.66	6.08	5.74	2.45	4.16	.	4.95	.	.	.	.	.	.	.
4	3.34	3.38	2.52	2.48	5.54	5.46	2.47	3.74	.	4.48	.	.	.	.	.	.	.
5	3.40	3.10	2.73	2.55	5.72	5.71	2.64	3.69	4.00	4.66	2.77	5.56	2.21	2.61	4.15	2.15	4.25
6	2.80	2.31	1.99	1.79	4.39	4.69	2.05	3.13	2.53	.	2.78	4.79	3.12	2.86	3.97	2.70	4.40
7	2.73	2.66	2.02	2.24	5.07	5.12	2.05	3.30	3.30	.	2.80	5.15	2.28	2.49	4.34	1.81	3.54
8	2.77	2.48	2.53	.	.	4.93	2.37	.	3.00	.	2.72	.	.	.	.	.	.
9	2.78	3.23	2.70	2.61	6.24	5.77	2.56	3.82	4.03	4.91	2.94	5.41	2.88	2.57	.	2.44	4.27
10	3.00	2.76	1.59	2.07	5.04	4.56	2.27	3.39	3.25	3.79	.	.	.	.	.	.	.

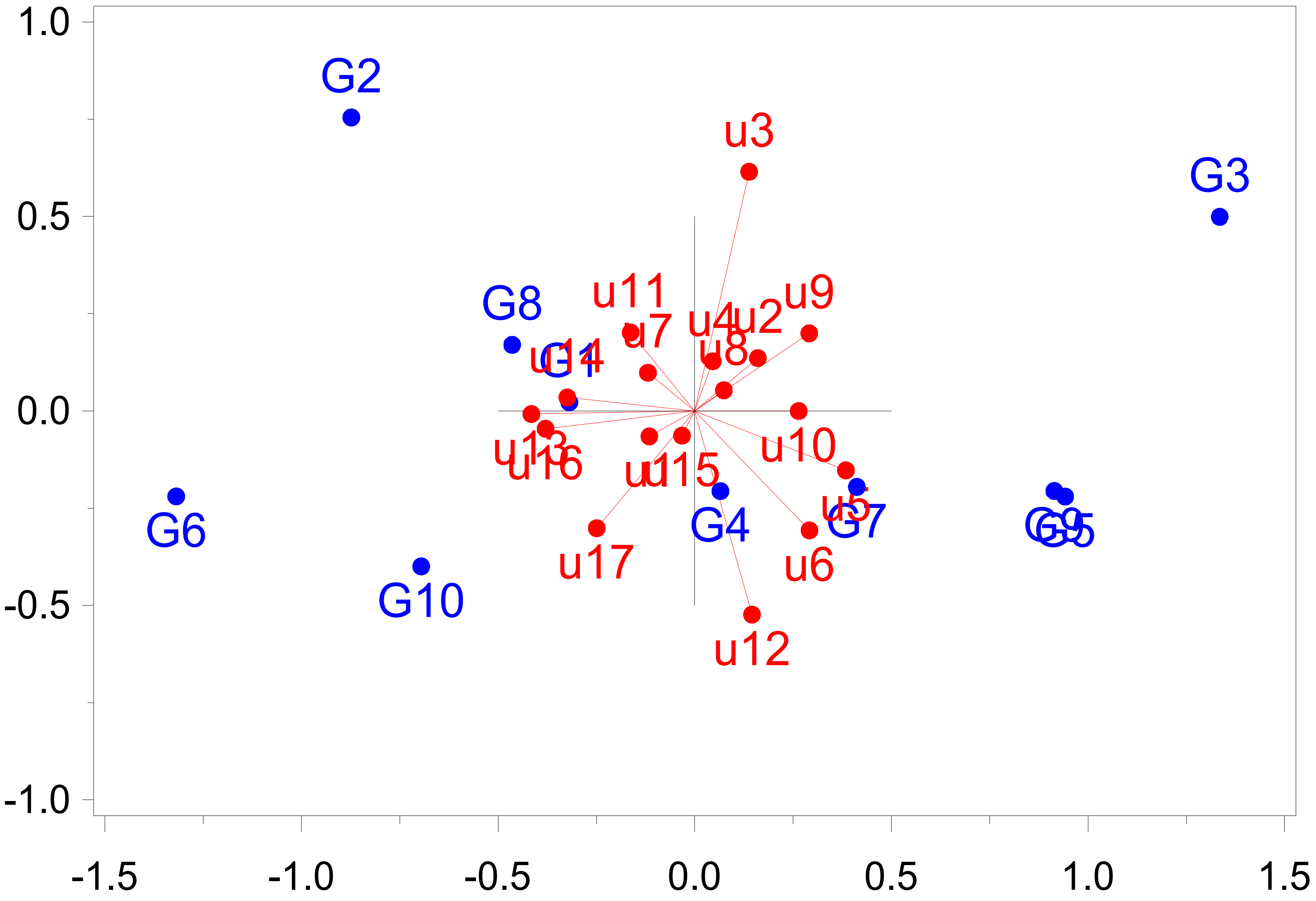


Fig. 3: Biplot for Dataset 2. Genotypes: G1-G10. Environments: u1-u17.

Tab. 6: Stability variances (Dataset 2).

Genotype	Stability variance	
G1	0.0931	
G2	0.1874	←
G3	0.1074	←
G4	0.0110	←
G5	0.0771	
G6	0.2274	←
G7	0.0239	
G8	0.0765	
G9	0.0865	
G10	0.0682	

Generation of biplots

- estimate interaction effects ⇒ BLUPs of $\lambda_{i1}w_{j1} + \lambda_{i2}w_{j2}$

- put BLUPs in matrix $\boldsymbol{W}$ (genotypes = rows and environments = columns)

- singular value decomposition (SVD) of the form $\boldsymbol{W} = \boldsymbol{U\Lambda V}$

  where $\boldsymbol{\Lambda}$ = diagonal matrix of singular values
  $\boldsymbol{U}$ and $\boldsymbol{V}$ = left and right singular vectors

- Scores $\boldsymbol{U\Lambda}$ and $\boldsymbol{V}$ ⇒ Euclidean distances between genotypes interpretable

6. Probability methods (Kent Eskridge, 1992)

Tab. 7: Hypothetical yields of 2 genotypes in 6 environments.

	Environment						Mean	Environ.
	A	B	C	D	E	F		variance
Genotype 1	20	27	21	25	22	23	23.0	6.0
Genotype 2	25	37	26	35	24	32	29.8	40.0

Example (Dataset 3)**:**

- CIMMYT data wheat
- South America, 1975 to 1986
- Very unbalanced: only 839 of 4500 year-location-genotype combinations
- 119 year-location combinations
- Square root transformation of data

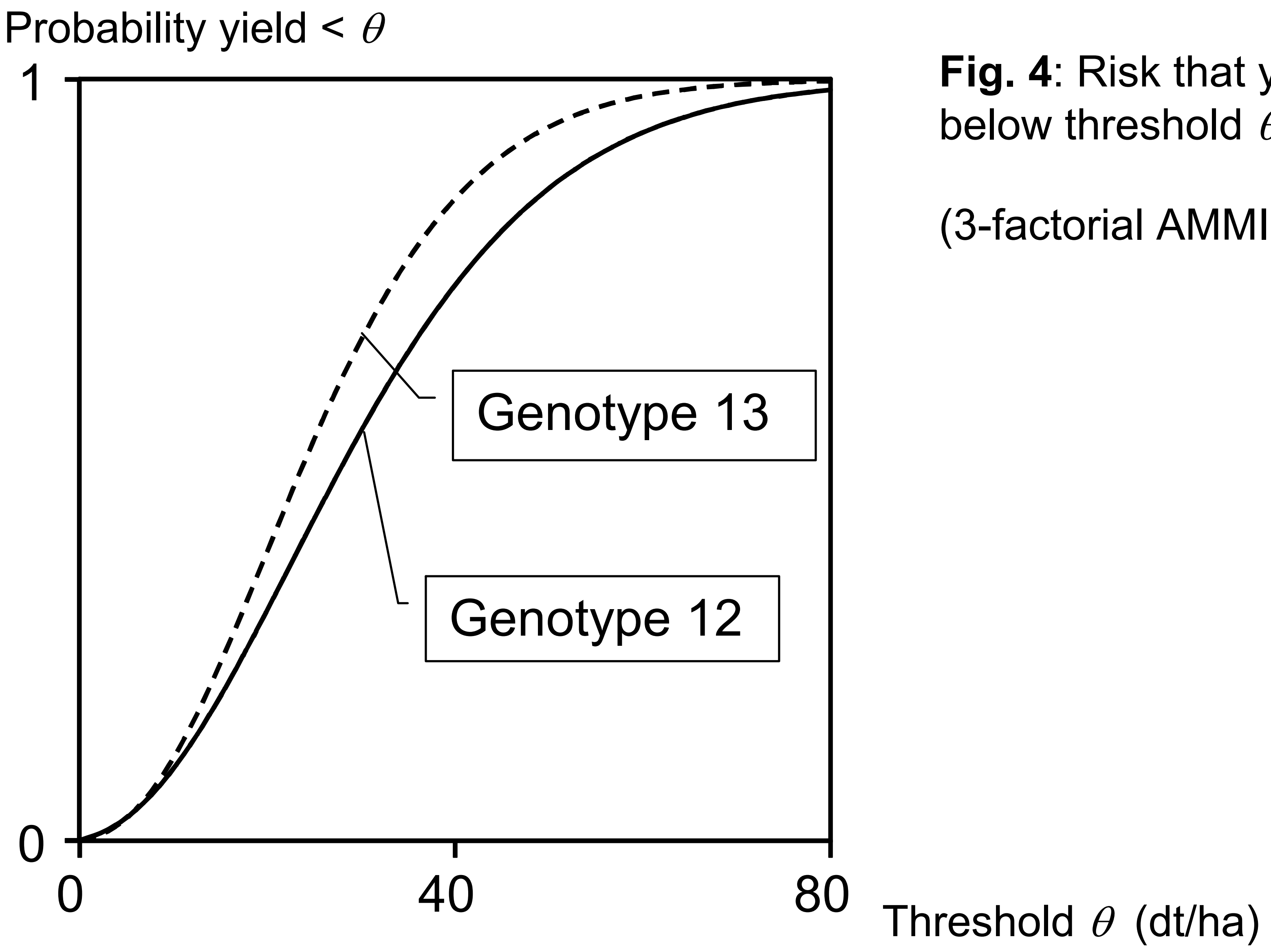


Fig. 4: Risk that yield falls below threshold θ (dt/ha).

(3-factorial AMMI model)

7. Three-factorial analysis of multi-year multi-location data

$y_{ijk} = \mu + g_i + f_{ij} + a_{ik} + b_{ijk}$

f_{ij} = effect of *i*-th genotype at *j*-th location

a_{ik} = effect of *i*-th genotype in *k*-th year

b_{ijk} = effect of *i*-th genotype at *j*-th location in *k*-th year

Find optimal variance-covariance structures for f_{ij}, a_{ik}, and b_{ijk}

Example:

f_{ij} = main effect *j*-th location + *ij*-th interaction genotype × location

a_{ik} = main effect *k*-th year + *ik*-th interaction genotype × year

b_{ijk} = main effect *jk*-th location-year combination

+ *ijk*-th interaction genotype - location - year

For each of the three interaction effects we can estimate a separate stability variance!

Example:

- CIMMYT data (wheat; Dataset 3)

f_{ij} = main effect *j*-th location + *ij*-th interaction genotype × location

⇒ ANOVA Modell

a_{ik} = main effect *k*-th year + *ik*-th interaction genotype × year

⇒ dropped (series spans a whole continent)

b_{ijk} = main effect *jk*-th location-year combination

+ *ijk*-th interaction genotype - location - year

⇒ AMMI1

Tab. 8: Simulated probability to outperform all other 14 genotypes (CIMMYT data; 1.000.000 simulation runs; Piepho & van Eeuwijk, 2002).

Genotype	Probability
1	0.0858
2	0.0385
3	0.0769
4	0.0585
5	0.0341
6	0.0253
7	0.0081
8	0.0389
9	0.0060
10	0.0517
11	0.0401
12	0.2318 ←
13	0.1038
14	0.1864
15	0.0141

8. Making use of environmental covariate information

Example:

- Triticale, regional yield trials (Germany)
- *Ackerzahl* (AZ) = soil fertility score (0-100)
- Regression approach
- 2 varieties: Alamo and Modus

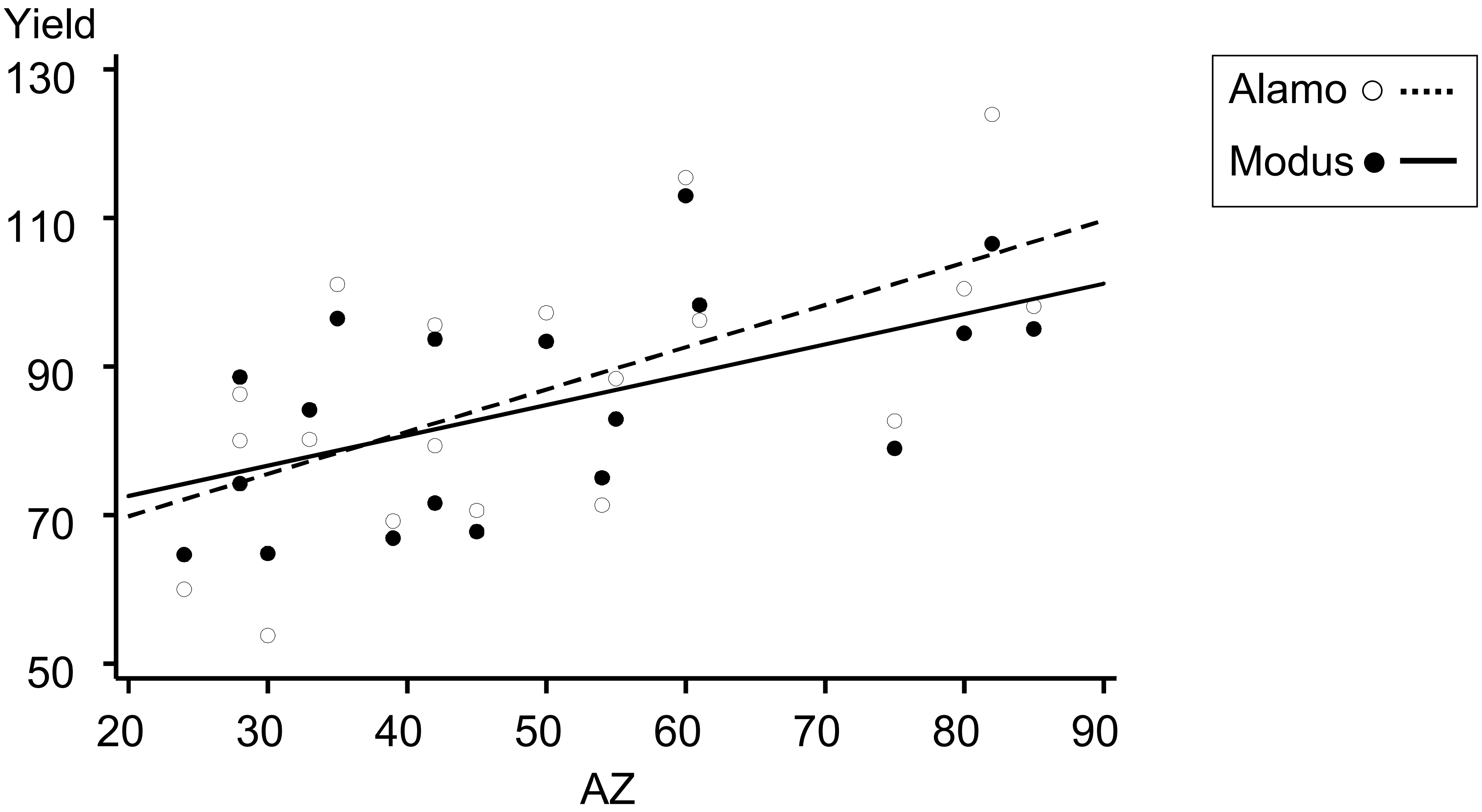


Fig. 6: Regression of yield on Ackerzahl (*AZ*) for 2 triticale varieties.

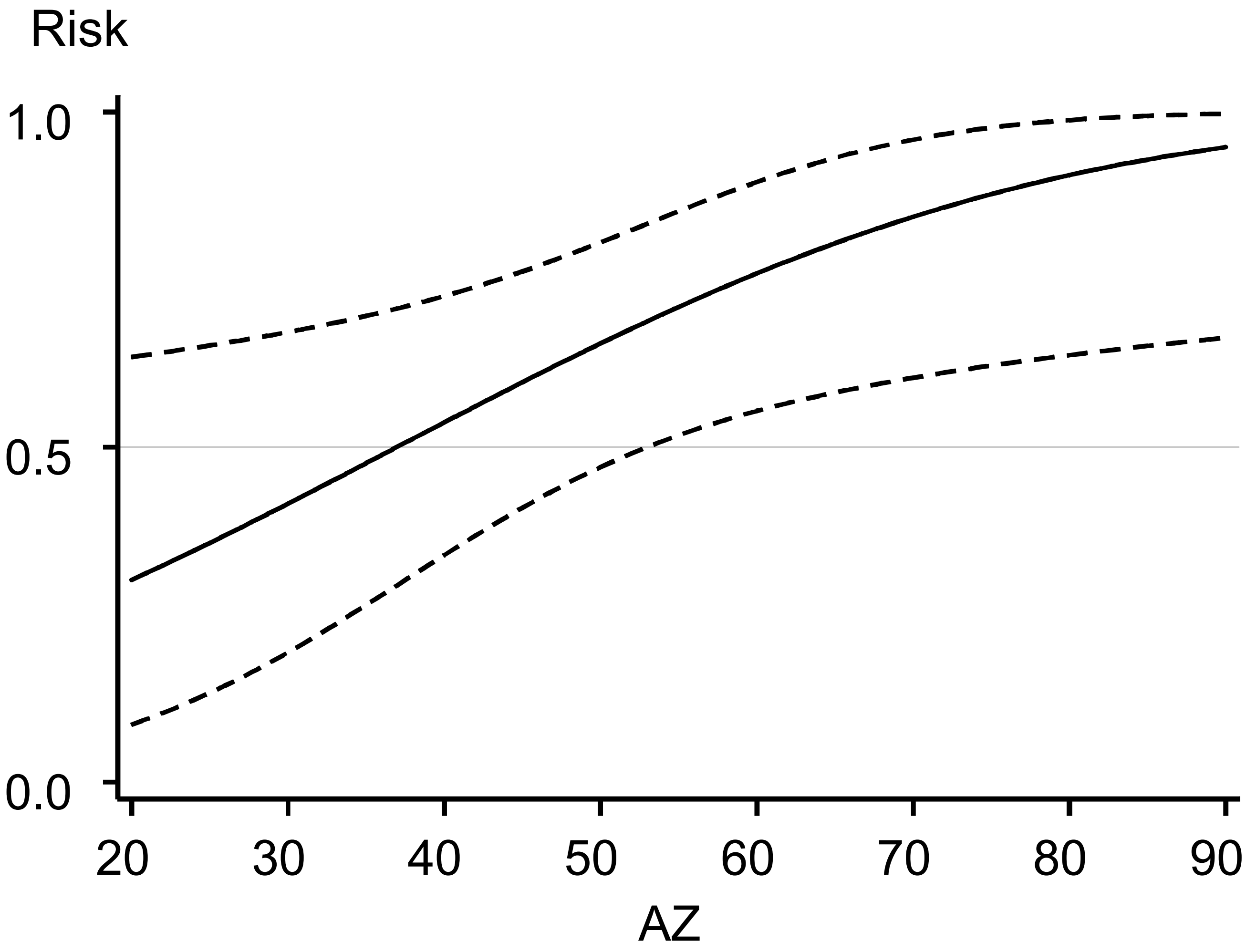
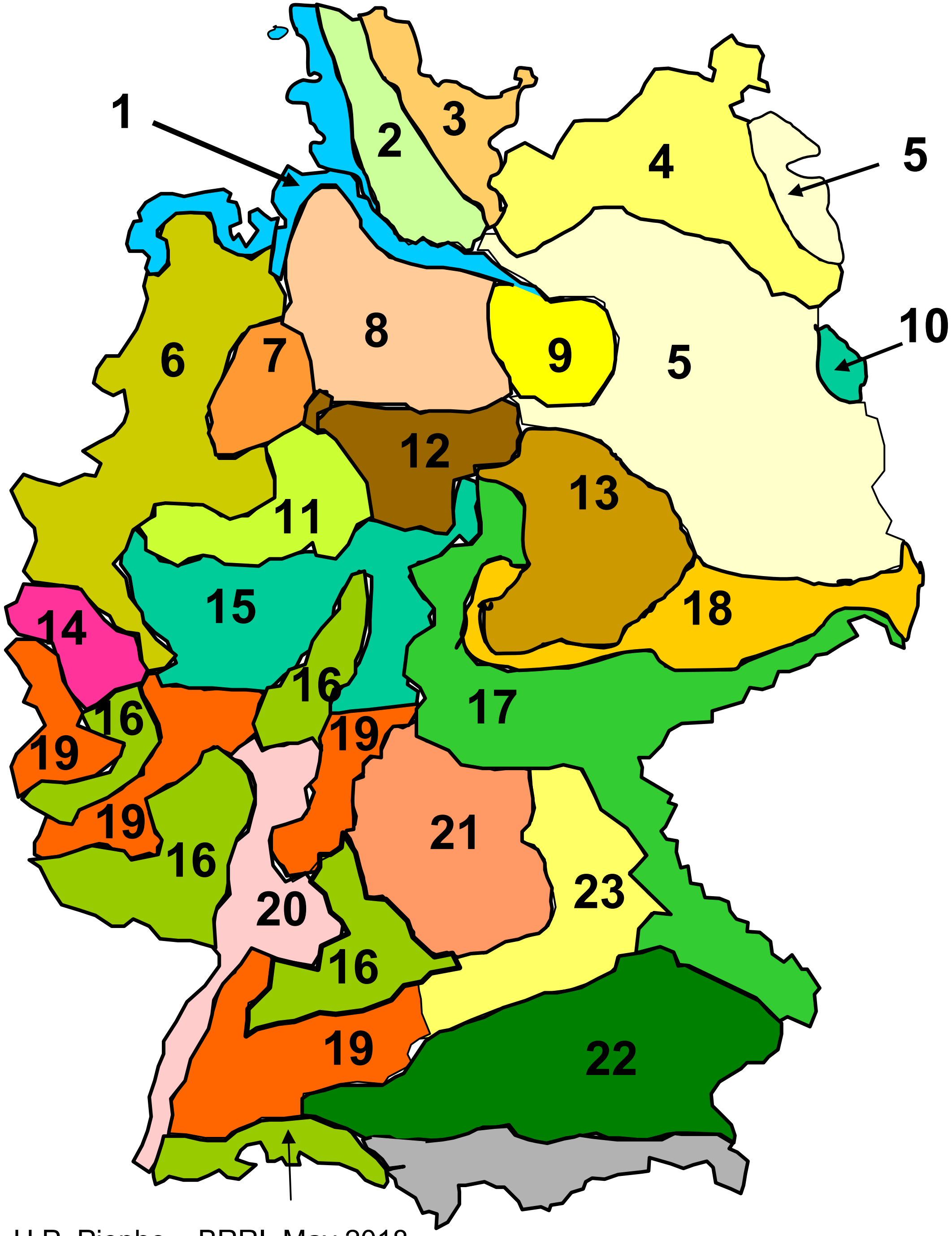


Fig. 7: Risk that Alamo is outperformed by Modus (——) as a function of *Ackerzahl* (*AZ*), with 95% confidence limits (·······). (Piepho, 2000)



Qualitative information

Example winter wheat:

- 23 agro-ecological zones in Germany

Two objectives

(1) **Local adaption**: Efficient estimation for spezific zone

⇒ Common approach: just use yield data per zone (inefficient)

(2) **Global adaption**: Efficient estimation of overall mean in target population of environments (across all zones)

⇒ Common approach: ignore stratification into zones

Improvement:

- Use all data for both objectives
- Use zones for both objectives
- Use BLUP of genotype-zone effects (Kleinknecht et al. 2013)

9. Time trend for stability

Does stability variance change over time?

Example from our trend analysis:

$$\operatorname{var}(GL)_{ij} = \sigma^2_{GL(1)} + r_i \sigma^2_{GL(2)}$$

$$\operatorname{var}(GY)_{ik} = \sigma^2_{GY(1)} + r_i \sigma^2_{GY(2)} + t_k \sigma^2_{GY(3)}$$

$$\operatorname{var}(GYL)_{ijk} = \sigma^2_{GYL(1)} + r_i \sigma^2_{GYL(2)} + t_k \sigma^2_{GYL(2)}$$

r_i = year of release of variety *i*

t_k = calendar year

⇒ These are linear regression lines for variance!

⇒ Allow variances to become negative: intercepts and slopes can be negative!

Coding the effects in a mixed model package

Genotype-location interaction:

$G \bullet L + G \bullet L \bullet \sqrt{r_i}$

Genotype-year interaction:

$G \bullet Y + G \bullet Y \bullet \sqrt{r_i} + G \bullet Y \bullet \sqrt{t_k}$

Genotype-year interaction:

$G \bullet L \bullet Y + G \bullet L \bullet Y \bullet \sqrt{r_i} + G \bullet Y \bullet L \bullet \sqrt{t_k}$

10. Concluding remarks

- Many stability measures can be defined and estimated as parameters of a mixed model (REML)
- This allows a unified view on stability analysis and different measures of stability
- Choice of stability measure boils down to covariance model selection problem
- There is often a variance - mean trade-off
- Probability methods and methods using environmental information particularly promising

References

Kleinknecht, K., Möhring, J., Singh, K.P., Zaidi, P.H., Atlin, G.N., Piepho, H.P., 2013: Comparison of the performance of BLUE and BLUP for zoned Indian maize data. *Crop Science* **53**, 1384-1391.

Piepho, H.P., 1998: Methods for comparing the yield stability of cropping systems - A review. *Journal of Agronomy and Crop Science* **180**, 193-213.

Piepho, H.P., 1999: Stability analysis using the SAS system. Agronomy Journal 91, 154-160.

Piepho, H.P., 2000: A mixed model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* **156**, 253-260.

Piepho, H.P., 2000: Exact confidence limits for covariate-dependent risk in cultivar trials. *Journal of Agricultural, Biological and Environmental Statistics* **5**, 202-213.

Piepho, H.P., Möhring, J., 2005: Best linear unbiased prediction for subdivided target regions. *Crop Science* **45**, 1151-1159.

Piepho, H.P., Möhring, J., 2006: Selection in cultivar trials – is it ignorable? *Crop Science* **146**, 193-202.

Piepho, H.P., van Eeuwijk, F.A., 2002: Stability analyses in crop performance evaluation. pp. 315-351. In: Kang, M. (ed): *Crop improvement: Challenges in the 21st century*. Haworth Press, New York.