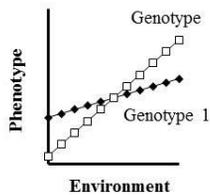


Estimation of GxE in animal breeding populations and implications of GxE for breeding programs

Han Mulder



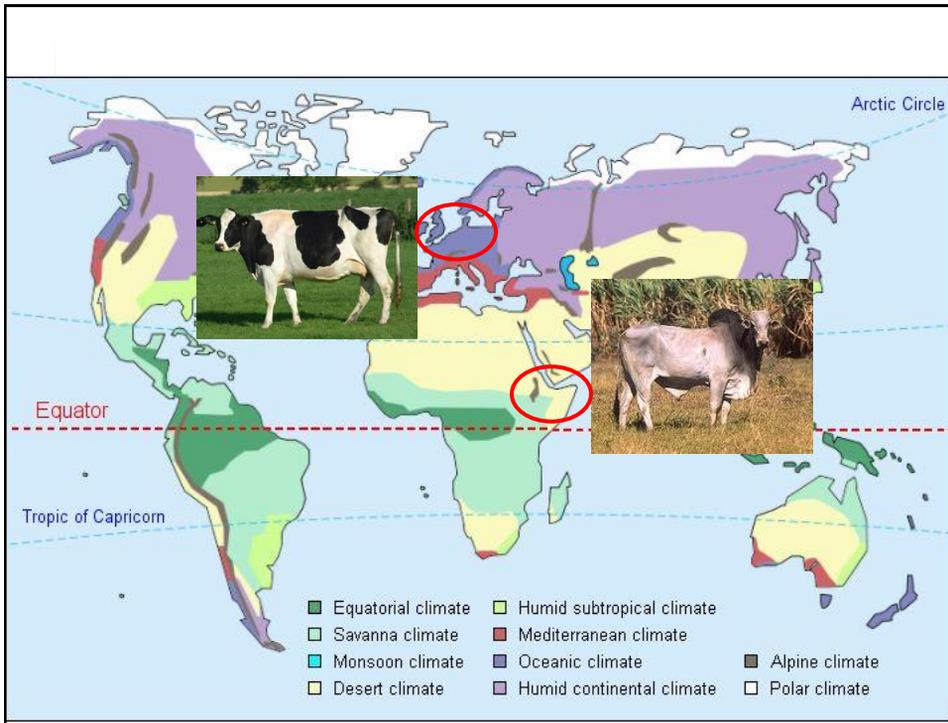
Contents

- Types of environments and size of GxE in animals
- Data structures to estimate GxE
- Dealing with GxE in breeding programs
- Statistical methods to estimate GxE in animal breeding
- Practical
 - Designs/data structures
 - Statistical analysis using ASReml

Learning outcomes

- To design experiments/datasets for estimating genetic correlations between environments
- To understand the effect of G x E on breeding programs
- To use bivariate and random regression models to analyze genotype by environment interaction

Types of environments and GxE found in animals



Environment can have many sights!

- Climate
- Housing system
- Nutrition
- Disease pressure
- Stocking density

....



Different types of environments

- Mega-environments
 - Different countries, different climate zones

- Macro-environments
 - Different climates within farms
 - Different farm types (organic vs conventional)

- Micro-environments
 - Each animal has a different environments
 - Some animals diseased; others not

Different types of environments

- Types of environments
 - Categorical
 - Farm types
 - Presence or absence of disease
 - → bivariate/multivariate model
 - Continuous
 - Temperature
 - Daylength
 - Rainfall
 - → Reaction norm model

How to quantify size of GxE?

- In animal breeding: aim is genetic improvement of populations by selection
 - GxE causing reranking has biggest impact
- The degree of GxE is judged by the genetic correlations between environments
 - How much is the genetic correlation deviating from 1.0?

How large is G x E in livestock?

- In dairy cattle (many studies)
 - Production: >0.8
 - Fertility/longevity: 0.5-1.0
- In pigs (fewer studies)
 - 0.5-1.0 between environments with stress and without stress (e.g. disease, heat stress)
 - 0.6-1.0 between farms with different health status
- In poultry (very few studies)
 - 0.6-1.0 between nucleus and commercial environments

How large is G x E in aquaculture?

- Extensive review
 - Different species
 - Different traits
 - Different environments: temperature, diet, location, rearing and stocking density

Table 5 Unweighted and weighted mean genetic correlations, number of observations (N), minimum (min) and maximum (max) by species and environments for growth traits

Species	Variable	Macro-environment [†]					Reference
		REAR	TEMP	DIET	DENT	LOCAT	
Nile tilapia	Mean [‡]	0.77 ⁸⁹	–	–	–	0.71 ⁹⁶	Elnath et al. (2007); Khaw et al. (2009); Thodesen et al. (2011); Bentzen et al. (2012); Khaw et al. (2012); Trong et al. (2013); Luan (2010); Luan et al. (2006)
	N	57	–	–	–	3	
	Min-max	0.07 to 0.99	–	–	–	0.45 to 0.87	
Tilapia shilensis	Mean	0.77 ⁸⁰	–	–	–	–	Makawa et al. (2006)
	N	3	–	–	–	–	
	Min-max	0.63 to 0.95	–	–	–	–	
Rainbow trout	Mean	0.47 ⁴⁶	0.36 ^{NA}	0.86 ⁹⁰	0.77 ^{NA}	–	McKay et al. (1984); Sylvén et al. (1991); Bagley et al. (1994); Kause et al. (2003, 2006); Pierce et al. (2008); Le Boucher et al. (2011a); Sae-Lim et al. (2013)
	N	25	2	23	10	–	
	min-max	0.15 to 0.86	0.18-0.54	0.55 to 1.00	0.56 to 0.90	–	
Atlantic cod	Mean	0.89 ⁹⁰	–	–	–	0.89 ⁸⁸	Kistad et al. (2006)
	N	2	–	–	–	2	
	Min-max	0.83 to 0.94	–	–	–	0.82 to 0.95	
Common carp	Mean	0.84 ⁸⁵	–	–	–	–	Ninh et al. (2011)
	N	3	–	–	–	–	
	Min-max	0.81 to 0.88	–	–	–	–	
European seabass	Mean	0.73 ⁸⁷	0.49 ^{NA}	0.78 ⁸⁹	0.51 ^{NA}	–	Sallant et al. (2006); Dupont-Nivet et al. (2008, 2010); Le Boucher et al. (2011b)
	N	18	1	7	1	–	
	Min-max	0.21 to 0.99	0.49	0.51 to 0.99	0.51	–	
Pacific white shrimp	Mean	0.87 ⁹³	–	–	0.84 ⁸⁹	–	Gitterle et al. (2005); Castillo-Juárez et al. (2007)
	N	14	–	–	3	–	
	Min-max	0.65 to 0.99	–	–	0.80-0.85	–	
Chinook Salmon	Mean	0.58 ^{NA}	–	–	–	–	Winkelman and Peterson (1994)
	N	6	–	–	–	–	
	Min-max	0.45 to 0.64	–	–	–	–	
Pacific oyster	Mean	0.74 ⁸⁹	–	–	–	0.81 ⁸⁴	Dégremont et al. (2007); Swan et al. (2007)
	N	16	–	–	–	27	
	Min-max	0.11 to 0.97	–	–	–	0.02 to 0.97	
Blue Mussel	Mean	–	–	–	–	0.58 ^{NA}	Mallet et al. (1986)
	N	–	–	–	–	1	
	Min-max	–	–	–	–	0.58	
European whitefish	Mean	–	–	0.97 ^{NA}	–	–	Quinton et al. (2007a)
	N	–	–	1	–	–	
	Min-max	–	–	0.97	–	–	
Common sole	Mean	0.42 ⁴⁰	–	–	–	–	Mas-Muñoz et al. (2013)
	N	2	–	–	–	–	
	Min-max	0.27 to 0.56	–	–	–	–	
Asian seabass	Mean	0.99 ^{NA}	–	–	–	–	Domingos et al. (2013)
	N	2	–	–	–	–	
	Min-max	0.98 to 0.99	–	–	–	–	
Red tilapia	Mean	0.72 ⁷⁸	–	–	–	–	Thodesen et al. (2013)
	N	7	–	–	–	–	
	Min-max	0.24 to 1.00	–	–	–	–	

Data structures to estimate G x E

Data structures to estimate G x E

- Categorical environments

- Measure genotype in different environments
 - Ideal design: animal itself or clones
 - Often animals perform in only one environment

- In animal breeding: no clones, no experiments!
 - Extensive databases with animal phenotypes and pedigree
 - High-density SNP-genotypes

How to estimate G x E?

- Usually pedigree links
 - Use of additive genetic relationships
 - E.g. half-sisters in different environments
 - Grand-offspring in different environments
 - E.g. less related individuals

- Use of genomic relationships
 - Example Silva et al. (2014; J. Anim. Sci. 92:3825-3834)

What kind of design is really needed?

- How much does the design affect the standard error on estimated genetic correlation?
 - How many families do we need with offspring in both environments?
 - N
 - How large should families be?
 - Number of offspring per environment: n
 - What is the effect of the heritability?
 - h^2

Standard error to estimate genetic correlation: Robertson (1959; Biometrics)

- Other formula

$$\text{se}(r_g) \approx \sqrt{\frac{[1+nt(1-r_g^2)]^2+r_g^2}{(N-1)n^2t^2}}$$

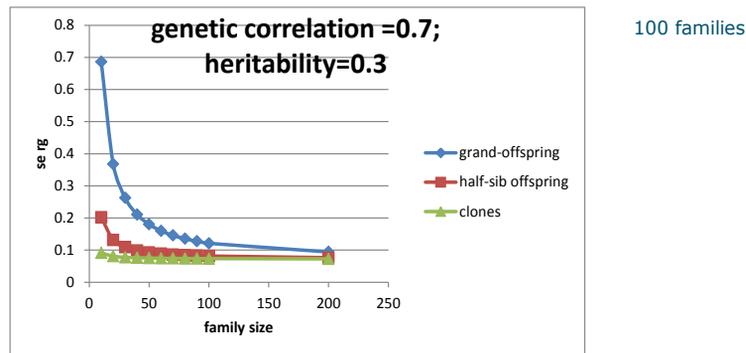
- t = intraclass correlation, e.g. $t = 0.25h^2$ for half-sibs

Bijma and Bastiaansen (2014, GSE)

$$\text{se}(r_g) \approx \sqrt{\frac{\frac{1}{r_{IH,x}^2 r_{IH,y}^2} + (1 + \frac{0.5}{r_{IH,x}^4} + \frac{0.5}{r_{IH,y}^4} - \frac{2}{r_{IH,x}^2 r_{IH,y}^2})r_g^2 + r_g^4}{(N-1)}}$$

- $r_{IH,x}^2$ = reliability of EBV in environment x = accuracy squared

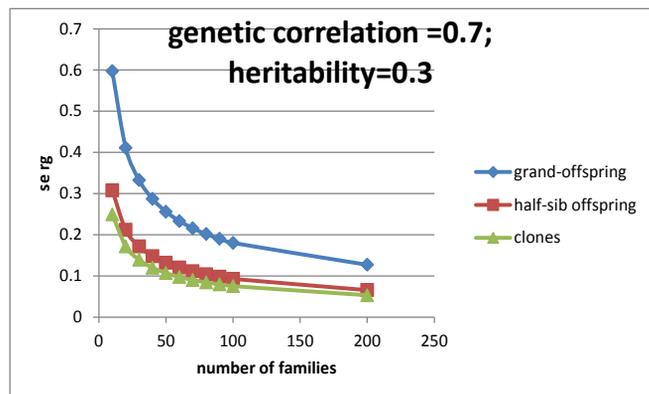
The effect of family size on $se(r_g)$



- Need many more grand-offspring than half-sib offspring/clones
- Clones is most efficient, but not feasible in livestock

The effect of number of families on $se(r_g)$

50 offspring
per family per
environment



Need 50-100 families to get accurate estimate of genetic correlation

Summary

- Large datasets required to estimate genetic correlations between environments
 - 50-100 families
 - Each with 50-100 offspring

- Clones slightly better than half-sibs, grand-offspring is quite a bit worse than half-sibs

Deal with GxE in breeding programs

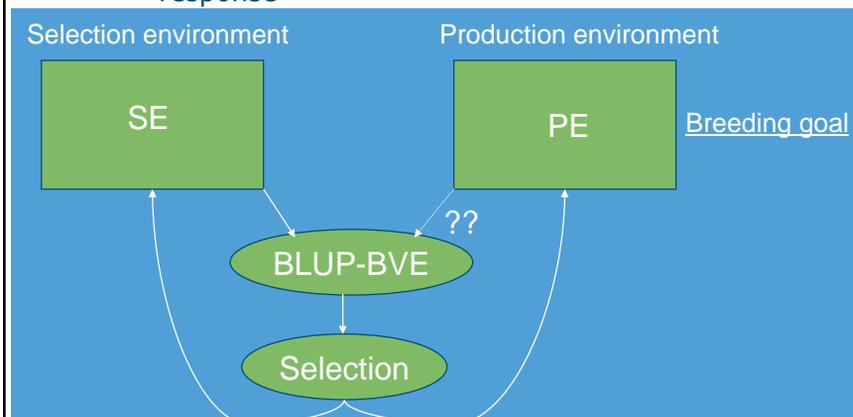
Different situations:

1. Nucleus and commercial environment

- Typically selection environment (SE) and production environment (PE) different
 - SE: higher health status, less diseases, optimal management
 - PE: higher disease pressure, lower management level, in pigs and poultry crossbred animals

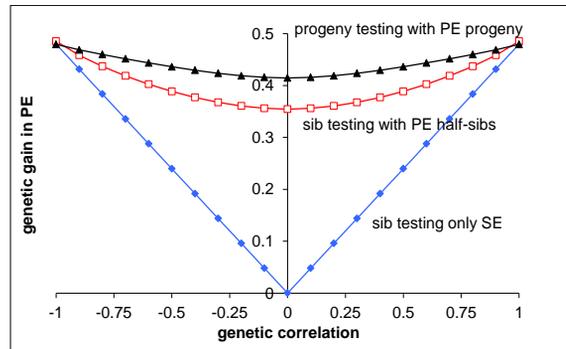
G x E: nucleus and production environment

- Only information from nucleus, but breeding goal is commercial environment
 - Genetic gain in commercial environment is correlated response



G x E: nucleus and production environment

- Use of sib/progeny information from commercial environment

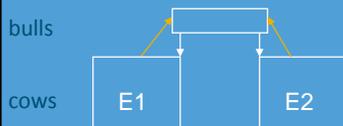


Different situations: 2. Multiple production environments

- Breeding organization are international
- Multiple production environments
 - Different climates
 - Within countries different types of farms
 - Organic and conventional
 - Management level
 - Barn type
 - Disease status
 - Grazing and non-grazing

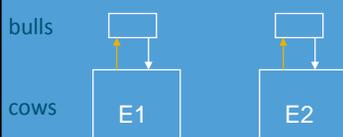
G x E: How many lines/breeding programs?

Breeding strategies



One breeding program

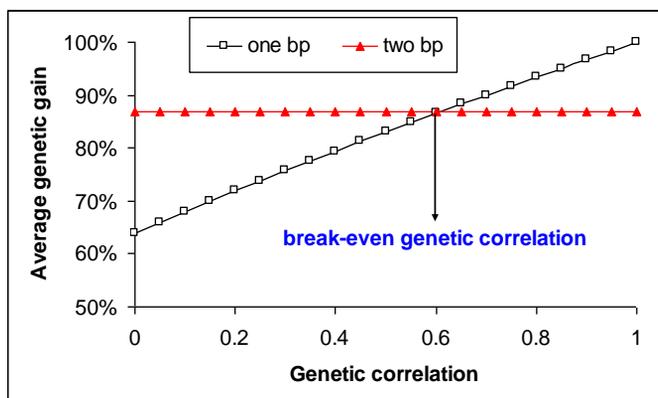
- all 400 bulls tested in both environments: 50 daughters in each environment
- increase average performance



Two breeding programs

- 200 bulls tested in one environment: 100 daughters in one environment
- Each bp: increase performance in environment of testing

G x E: How many lines/breeding programs?



G x E and multi-trait selection

- Between environments
 - G x E per trait
 - Heterogeneity of genetic variances
 - Breeding goal differences
 - Different genetic correlations between traits
 - **Genetic correlation between breeding goals**
 - $$r_{H,kl} = \frac{\mathbf{v}_k' \mathbf{G} \mathbf{v}_l}{\sqrt{\mathbf{v}_k' \mathbf{G} \mathbf{v}_k \mathbf{v}_l' \mathbf{G} \mathbf{v}_l}}$$
 - **G**: full genetic variance-covariance matrix between all traits in the breeding goals of environment k and l
 - \mathbf{v}_k : economic values for environment k
 - \mathbf{v}_l : economic values for environment l
- (Mulder, 2007)

Summary

- G x E lowers genetic gain, but more genetic diversity is conserved
- Nucleus and production environment
 - Minimize environmental difference
 - Use phenotypes of sibs or progeny in multivariate breeding value estimation
- Different production environments
 - If $rg > 0.6-0.7$ then single breeding program (provided that information of sibs/progeny is collected in both environments)
 - If $rg < 0.6-0.7$, then different breeding programs needed

Statistical methods to estimate GxE in animal breeding

Statistical methods to estimate G x E in animal breeding populations

- We use pedigree relationships and we use BLUP
- Main interest in additive genetic effects or breeding values
- Most common models to analyze G x E
 - Bivariate/multivariate models
 - Reaction norm/random regression models

BLUP

- $y = \mu + herd + animal + e$
- y = phenotype
- μ = fixed mean
- $herd$ = fixed effect for herd
- $animal$ = random additive genetic effect = EBV
- e = residual



BLUP mixed model equations

- $y = Xb + Za + e$
- X =design matrix to link phenotypes to fixed effects, e.g. which cow is in which herd
- b =vector with solutions for fixed effects
- Z =design matrix to link phenotypes to EBV
- a =vector with EBV for all animals
- $$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$
- $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$
- A^{-1} =inverse of additive genetic relationship matrix

Breeding values Holstein bulls

	kg milk	%fat	%protein	kg fat	kg protein	total merit index milk	total merit index NVI
Bookem	+1552	-0.29	-0.11	+37	+43	+275	+299
G-Force	+771	+0.16	+0.13	+48	+39	+271	+246
Atlantic	+298	-0.06	+0.14	+7	+23	+110	+245
Titanium	+645	+0.21	+0.01	+47	+23	+199	+234
Snowman	+2576	-0.37	-0.29	+69	+57	+415	+228



Bivariate model to estimate GxE

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & \mathbf{0} \\ \mathbf{0} & Z_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \sim N \left[\begin{bmatrix} \mathbf{0} \\ \mathbf{0}' \end{bmatrix}, A \otimes \begin{bmatrix} \sigma_{a1}^2 & \sigma_{a1,a2} \\ \text{symmetric} & \sigma_{a2}^2 \end{bmatrix} \right]$$

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \sim N \left[\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0} \end{bmatrix}, \begin{bmatrix} I_1 \sigma_{e1}^2 & \mathbf{0} \\ \mathbf{0} & I_2 \sigma_{e2}^2 \end{bmatrix} \right]$$

- no residual covariances between environments if animals are in one environment

$$r_g = r_a = \frac{\sigma_{a1,a2}}{\sigma_{a1}\sigma_{a2}}$$



Reaction norm models

- $y = \text{fixed effects} + bx + a_{int} + a_{sl}x + e$
- Fixed reaction norm: bx
- $\begin{bmatrix} \mathbf{a}_{int} \\ \mathbf{a}_{sl} \end{bmatrix} \sim N \left[\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{A} \otimes \begin{bmatrix} \sigma_{aint}^2 & \sigma_{aint,asl} \\ \text{symmetric} & \sigma_{asl}^2 \end{bmatrix} \right]$
- \mathbf{A} = matrix with all additive genetic relationships
- Heterogeneity of residual variance accounted for using 3-10 groups each with their own residual variance



Issues with reaction norms models

1. Which covariate to use?
 1. External environmental factor
 2. Internal data derived parameter
2. Scaling and (Legendre) polynomials
3. Model comparison
4. Interpretation of results



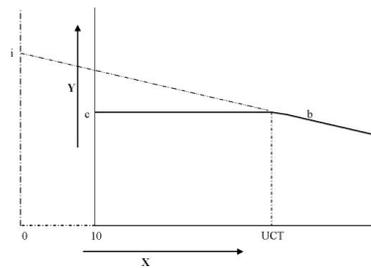
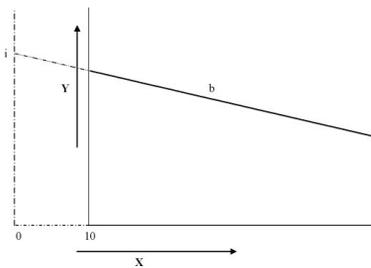
1. Which covariate to use?

External and internal environmental factors

- External environmental factors
 - Temperature
 - Day length
 - Rainfall
 - Salinity, oxygen (fish)
 - ...
- Internal derived environmental factors
 - Finlay-Wilkinson regression
 - Mean performance
 - Herd-year-season estimated effect

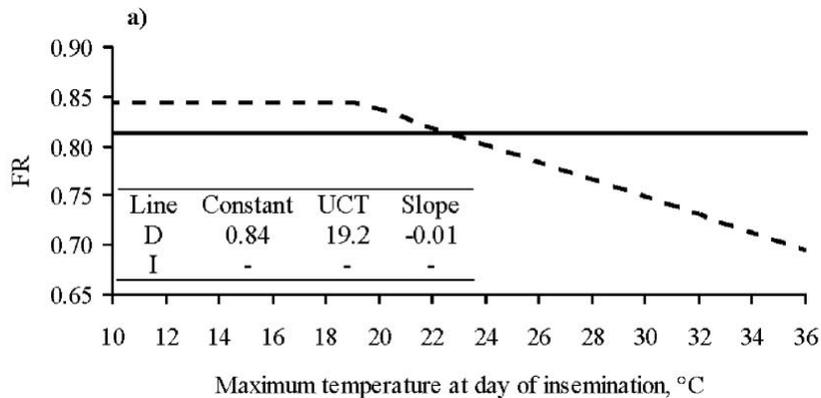
Temperature

- Temperature may affect phenotypes only above a certain temperature, the so-called upper critical temperature



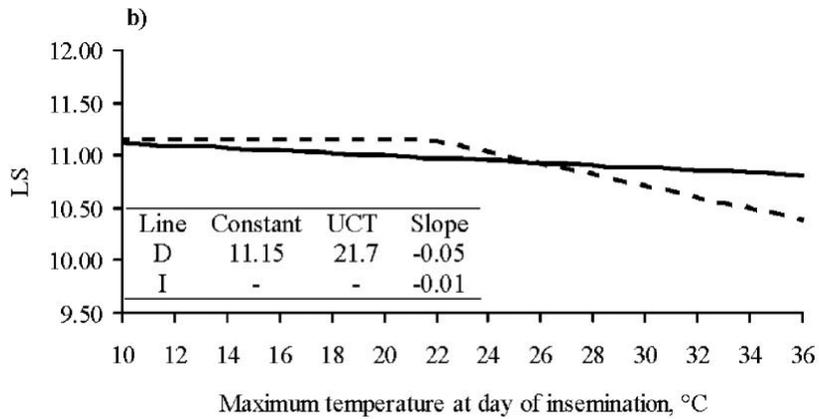
Bloemhof et al., 2008. J. Anim. Sci. 86:3330-3337

Upper critical temperature: Farrowing rate at first insemination



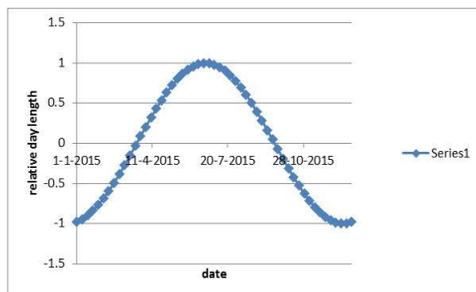
Bloemhof et al., 2008. J. Anim. Sci. 86:3330-3337

Upper critical temperature: litter size



Day length

- Day length is according to a sinus function



Internal data derived parameter: The use of herd-year-season estimates

- In many studies, it is difficult to categorize farms
 - No access to external data
 - HYS gives indication of management level, but may also contain the genetic level of the herd

- Strategy
 - Estimated herd-year-season effects on the same data using mixed model
 - Add to the data set and use random regression
 - Use of data twice = tricky



Possible solutions

- Derive environmental parameters (EP) from other traits
 - Calus et al. (2003; J. Dairy Sci. 86: 3756-3764)

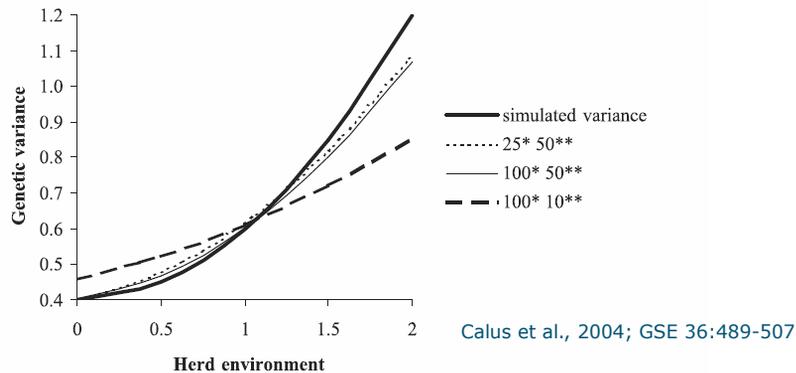
- Use other animals to calculate the EP

- Use many animals to estimate EP, dependency is smaller
 - Avoid very small HYS classes
 - Include all parities



Consequences for estimation of G x E

- Reaction norm models tend to underestimate G x E
 - Underestimation of the genetic variance in slope
 - Correlations closer to 1.0 than the true value



Possible solution

- Bayesian approach
 - The x-variable is simultaneously sampled with the breeding values and the other effects in the model

Table 1. Mean and SE of estimates (based on posterior means) of (co)variance components over 20 replicate simulations

Model ¹	$\sigma_{a_0}^2$	$\sigma_{a_h}^2$	$\sigma_{a_0 a_h}$	σ_e^2
Realized ²	100.4 ± 0.040	1.01 ± 0.002	5.11 ± 0.065	298.3 ± 0.016
M1 ³	101.7 ± 1.102	1.02 ± 0.034	5.04 ± 0.101	297.1 ± 0.872
M2 ⁴	99.3 ± 1.051	1.01 ± 0.013	5.00 ± 0.080	298.5 ± 0.868
M3 ⁵	111.5 ± 1.440	0.58 ± 0.020	3.68 ± 0.105	305.5 ± 0.702

¹ $\sigma_{a_0}^2$ = variance of the level; $\sigma_{a_h}^2$ = variance of the slope of additive genetic reaction norm; $\sigma_{a_0 a_h}$ = covariance between the level and the slope; and σ_e^2 = residual variance.

²The variance components were calculated from the realized values of the simulation.

³Model with unknown covariate of reaction norm (the proposed approach).

⁴Model using true herd-year effect as covariate of reaction norm.

⁵Model using phenotypic mean of herd-year as covariate of reaction norm.

Su et al., J. Anim. Sci.
84:1651-1657

2. Scaling and (Legendre) polynomials

Do we need polynomials?

- Linear reaction norm
 - No need for use of polynomials
 - Would give equivalent results

- Higher order reactions norms
 - Yes, performance of REML or Gibbs much better

Linear reaction norm models without polynomials

- Scaling of covariate mean = 0, variance 1.0
 - The correlation between intercept and slope has a meaning when selection is performed in the average environment
 - Variance of 1.0 makes it feasible to compare estimates of genetic variance in slope when using different covariates

Higher reaction norms: Legendre polynomials

- They are orthogonal
 - Lower correlations between regression coefficients
– faster convergence
- Scale the EP to be between -1 and 1
- $$x_l = -1 + 2 * \left(\frac{EP_l - EP_{min}}{EP_{max} - EP_{min}} \right)$$

Legendre polynomials

- Legendre polynomial coefficient order $n > 1$, recursive equation:

- $P_0 = 1$

- $P_1 = x$

- $$P_{n+1}(x) = \frac{1}{n+1}((2n+1)xP_n(x) - nP_{n-1}(x))$$

- $$\phi_n(x) = \left(\frac{2n+1}{2}\right)^{0.5} P_n(x)$$



(Schaeffer, Random regression models:
<http://www.aps.uoguelph.ca/~lrs/ABModels/NOTES/RRM14a.pdf>)

Example polynomial coefficients

x	x scaled	P0	P1	P2	ϕ_0	ϕ_1	ϕ_2
100	0.33	1.00	0.33	-0.33	0.71	0.41	-0.53
200	0.67	1.00	0.67	0.17	0.71	0.82	0.26
300	1.00	1.00	1.00	1.00	0.71	1.22	1.58
-100	-0.33	1.00	-0.33	-0.33	0.71	-0.41	-0.53
-200	-0.67	1.00	-0.67	0.17	0.71	-0.82	0.26
-300	-1.00	1.00	-1.00	1.00	0.71	-1.22	1.58
0	0.00	1.00	0.00	-0.50	0.71	0.00	-0.79



3. Model comparison

Significance of model

- Likelihood ratio test

- H0: model with only intercept
- H1: model with intercept and slope

- The likelihood ratio:
 - $D = 2\log L(\text{full model}) - 2\log L(\text{reduced model})$
 - If the hypothesis contains a parameter on the boundary, then D follows a mixture of Chi-square distributions

Which degrees of freedom?

- Suppose the model under H0 estimates:

$$\mathbf{G} = \begin{bmatrix} \sigma_{aint}^2 & 0 \\ 0 & 0 \end{bmatrix}$$

- The model under H1:

$$\mathbf{G} = \begin{bmatrix} \sigma_{aint}^2 & \sigma_{aint,asl} \\ \text{symmetric} & \sigma_{asl}^2 \end{bmatrix}$$

- The large sample distribution is:
- Mixture of χ_1^2 and χ_2^2



Visscher, 2006; Twin research and human studies 9: 490-495
Stram and Lee, 1994; Biometrics 50:1171-1177

In more general terms

- Model H0:

$$\mathbf{G} = \begin{bmatrix} \mathbf{D}_0 & 0 \\ 0 & 0 \end{bmatrix}$$

- \mathbf{D}_0 is a matrix with dimension $q * q$

- Model H1:

$$\mathbf{G} = \mathbf{D}_1 = \begin{bmatrix} \mathbf{D}_0 & d12 \\ d21 & d22 \end{bmatrix}$$

- \mathbf{D}_1 is a matrix with dimension $(q+1) * (q+1)$

- The large sample distribution is:

- Mixture of χ_q^2 and χ_{q+1}^2



Stram and Lee, 1994; Biometrics 50:1171-1177

Other model comparisons

- Akaike's information criterion:
 - $AIC = -2\text{Log}L + 2t$
 - t = number of variance parameters in the model

- Bayesian information criterion (more conservative):
 - $BIC = -2\text{Log}L + 2t\log(v)$
 - v = residual degrees of freedom

- AIC/BIC are not tests for significance
- AIC/BIC favour the most parsimonious model

Other model comparisons

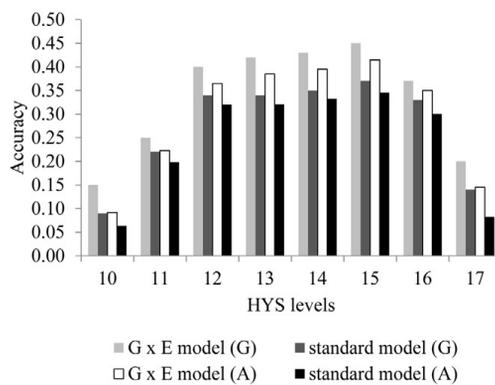
- Check the genetic parameters obtained from reaction norm model with a bivariate model

- Reaction norm models may lead to:
 - Extreme heritabilities in extreme environments
 - Low genetic correlation between extreme environments

Other model comparisons

- Predictive ability
- Cross-validation
 - Predict the phenotype or adjusted phenotype in the validation set

Accuracy of genomic and pedigree breeding values

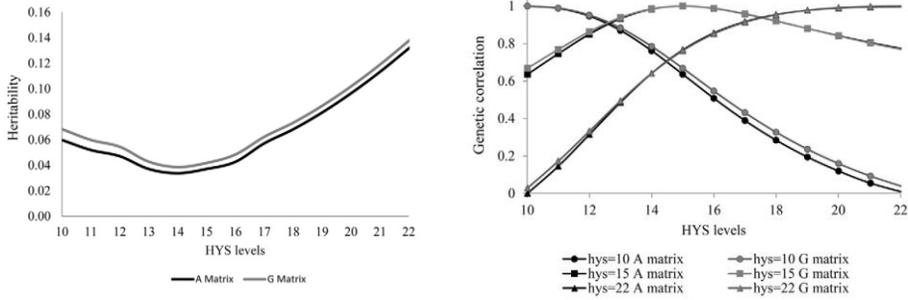


4. Interpretation of results

Calculation of genetic parameters

- Genetic variance-covariance matrix between different environments
- $\mathbf{H} = \Phi \mathbf{G} \Phi'$
- \mathbf{G} = matrix estimated (co)variances for the different orders of the polynomial
- Φ = matrix with ϕ values for the orders of the polynomial for the environments of interest.

Genetic parameters using genomic or pedigree relationship matrix (litter size)

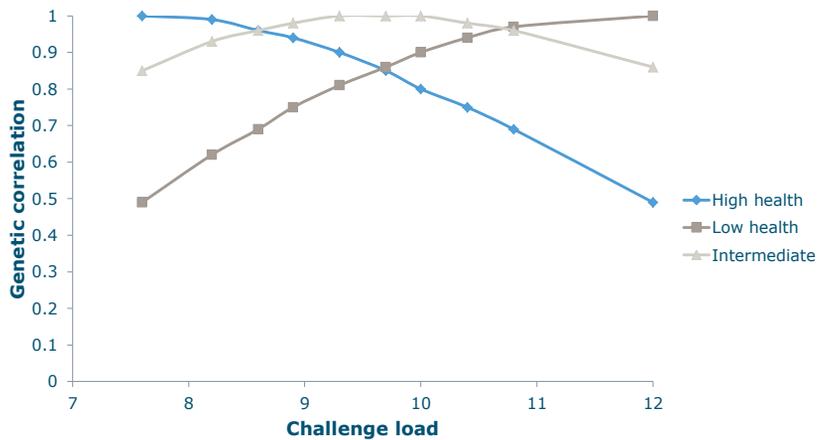


Trait: litter size
 Environments: Large White sows in 22 countries



Silva et al., 2014; J. Anim. Sci. 92:3825–3834

Genetic correlations between different environments



Herrero-Medrano et al., 2015; J. Anim. Sci. 93:1494-1502

Running ASREML with reaction norms

- ASREML mean model iteration1
- animal !P
- sire
- dam
- herd
- hys
- Am
- Asl
- Av
- x #!-1 # you can shift the intercept if you want
- E
- Pheno

- ped.dat !make
- cows_asreml.dat !MAXIT 100

- Pheno ~ mu !r animal animal.x
- 1 1 1
- 10000
- animal 2
- 2 0 US 0.3 0.0 0.05
- animal

Running ASReml

- Use ASREML-W

- Or with a batch-file

Summary

- Bivariate models and reaction norm models can be used to estimate $G \times E$
- Reaction norm models are more complex
 - Heritabilities and genetic correlations for every set or pair of environments
- Different environmental parameters can be used
 - Be careful when using HYS