Contents lists available at ScienceDirect

Animal
The international journal of animal biosciences

Estimation of macro- and micro-genetic environmental sensitivity in unbalanced datasets

M.D. Madsen [a,*], J. van der Werf [a], V. Börner [b,c], H.A. Mulder [d], S. Clark [a]

[a] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia
[b] Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia
[c] Centre for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark
[d] Animal Breeding and Genomics Centre, Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands

ARTICLE INFO

Article history:
Received 1 July 2021
Revised 21 October 2021
Accepted 22 October 2021
Available online 24 November 2021

Keywords:
Data structure
Double hierarchical generalised linear models
Genotype-by-environment interactions
Reaction norm
Simulation study

ABSTRACT

Genotype-by-environment interaction is caused by variation in genetic environmental sensitivity (**GES**), which can be subdivided into macro- and micro-GES. Macro-GES is genetic sensitivity to macro-environments (definable environments often shared by groups of animals), while micro-GES is genetic sensitivity to micro-environments (individual environments). A combined reaction norm and double hierarchical generalised linear model (**RN-DHGLM**) allows for simultaneous estimation of base genetic, macro- and micro-GES effects. The accuracy of variance components estimated using a RN-DHGLM has been explicitly studied for balanced data and recommendation of a data size with a minimum of 100 sires with at least 100 offspring each have been made. In the current study, the data size (numbers of sires and progeny) and structure requirements of the RN-DHGLM were investigated for two types of unbalanced datasets. Both datasets had a variable number of offspring per sire, but one dataset also had a variable number of offspring within macro-environments. The accuracy and bias of the estimated macro- and micro-GES effects and the estimated breeding values (**EBV**s) obtained using the RN-DHGLM depended on the data size. Reasonably accurate and unbiased estimates were obtained with data containing 500 sires with 20 offspring or 100 sires with 50 offspring, regardless of the data structure. Variable progeny group sizes, alone or in combination with an unequal number of offspring within macro-environments, had little impact on the dispersion of the EBVs or the bias and accuracy of variance component estimation, but resulted in lower accuracies of the EBVs. Compared to genetic correlations of zero, a genetic correlation of 0.5 between base genetic, macro- and micro-GES components resulted in a slight decrease in the percentage of replicates that converged out of 100 replicates, but had no effect on the dispersion and accuracy of variance component estimation or the dispersion of the EBVs. The results show that it is possible to apply the RN-DHGLM to unbalanced datasets to obtain estimates of variance due to macro- and micro-GES. Furthermore, the levels of accuracy and bias of variance estimates when analysing macro- and micro-GES simultaneously are determined by average family size, with limited impact from variability in family size and/or cohort size. This creates opportunities for the use of field data from populations with unbalanced data structures when estimating macro- and micro-GES.

Implications

Uniformity and sensitivity of genotypes to environmental variation are of increasing importance to livestock producers. Genotype-by-environment interactions, both within and across environments, can negatively impact uniformity. Datasets to estimate these interactions are usually unbalanced, for instance, having uneven family and/or cohort sizes. This study explored data structure requirements and shows that average family size, rather than variability in family size and/or cohort size, determines the levels of accuracy and bias when analysing genotype-by-environment interactions simultaneously within and across environments.

Introduction

A population under selection often contains animals exposed to different macro- and micro-environments. Macro-environments are definable environments such as location, herd, or contemporary group and may be shared by multiple animals, while

[*] Corresponding author.
E-mail address: mmadsen3@myune.edu.au (M.D. Madsen).

https://doi.org/10.1016/j.animal.2021.100411
1751-7311/© 2021 The Authors. Published by Elsevier B.V. on behalf of The Animal Consortium.

micro-environments are the part of the environmental effect experienced by an individual animal (Hill and Mulder, 2010). Both macro- and micro-environments influence the phenotype of an animal, depending in part on the animal's genetic environmental sensitivity (**GES**). Genetic environmental sensitivity is caused by the genetic constitution of the individual, that is to say individuals from different genetic backgrounds, e.g. different sire lines, may respond differently to environmental changes. This causes a variation in GES known as genotype-by-environment interaction (**G × E**). The consequence of GES depends on whether it is caused by macro- or micro-environmental changes and, therefore, models accounting for GES may contrast macro- and micro-GES. Macro-GES may cause reranking of animals, where the best genotype in one macro-environment is not the best in another, or can result in a change in scale across macro-environments, i.e. the difference between genotypes changes across macro-environments (Falconer and Mackay, 1996). Micro-GES affects the variability of phenotypes, e.g. offspring of a sire with low micro-GES will be less variable in phenotype than offspring of a sire with high micro-GES when environmental factors are the same for the offspring of both sires (Rönnegård et al., 2013).

Statistically, the impacts of macro- and micro-environments are modelled differently. If genotypes are measured across a wide range of macro-environments, G × E due to macro-GES can be studied using reaction norm models or by considering a trait expressed in different environments as different traits (Falconer and Mackay, 1996). G × E due to micro-GES is modelled as a genetic component affecting heterogeneity of the environmental variance, because micro-GES affects the variability of the phenotype. Estimation of the variance component associated with micro-GES has been based on using double hierarchical generalised linear models (**DHGLMs**) (Rönnegård et al., 2010; Felleki et al., 2012). The statistical framework behind DHGLMs was developed by Lee and Nelder (2006) with a focus on economic modelling. Rönnegård et al. (2010) adapted the DHGLM for the estimation of micro-GES under the assumption that the genetic effect due to micro-GES follows the exponential model described by SanCristobal-Gaudy et al. (1998). The DHGLM proposed by Rönnegård et al. (2010) iterates between a linear mixed model, termed the mean part, and a Gamma GLM, termed the dispersion part. Each part of the model contains a random genetic effect. Felleki et al. (2012) extended the model to estimate the genetic correlation between the genetic effects on the mean and those on the dispersion part and they also showed that linearising the dispersion phenotype around its current fitted value was equivalent to using a Gamma GLM. Mulder et al. (2013) included a linear reaction norm in the mean part of the DHGLM, such that the combined model (here termed **RN-DHGLM**) allows for simultaneous estimation of macro- and micro-GES. Mulder et al. (2013) showed that a suitable data structure for the RN-DHGLM included at least 100 sires with at least 100 offspring each, evenly distributed across macro-environments. It was shown that the estimates of genetic variance of micro-GES had unacceptably large SD across 100 replicates and the covariance matrix has to be forced to be positive definite in 8–51% of the replicates if the number of sires or number of offspring per sire was less than 100 (Mulder et al., 2013). However, field data are usually unbalanced in every sense, yet the data requirements of the RN-DHGLM in the case of unbalanced data have not been formally studied.

The aim of this study was thus to investigate the data requirements when estimating variance components and breeding values for macro- and micro-GES in unbalanced data using a RN-DHGLM.

Material and methods

Multiple data sizes and structures were simulated in order to investigate the impact of unbalanced data on the estimation of variance components and EBVs obtained using the RN-DHGLM.

Simulated phenotypes

In all simulations, phenotypes were generated following the model described by Mulder et al. (2013). The model included a base genetic effect and macro-GES to the effect of a macro-environment, expressed as the intercept and slope of a linear reaction norm, and micro-GES expressed as genetic heterogeneity of environmental variance following an exponential model (SanCristobal-Gaudy et al., 1998; Mulder et al., 2013).

$$y_{ij} = \mu + a_{inti} + a_{sli}x_j + exp\left(0.5ln\left(\sigma^2_{e_d}\right) + 0.5a_{di}\right)\varepsilon_i \tag{1}$$

where y_{ij} is the phenotype of animal i reared in macro-environment j, μ is the population mean, a_{inti}, a_{sli} and a_{di} are the breeding values for the intercept (base genetic effect), slope (macro-GES effect) and dispersion (micro-GES effect), respectively, for animal i, x_j is the random effect of macro-environment j, $\sigma^2_{e_d}$ is the environmental variance of the exponential model and ε_i is a random environmental variable drawn from $N(0,1)$. The breeding values a_{inti}, a_{sli} and a_{di} were drawn from $MVN(\mathbf{0},\mathbf{G}\otimes\mathbf{A})$, where **A** is the additive genetic relationship matrix derived from the pedigree and $\mathbf{G} = \begin{bmatrix} \sigma^2_{a_{int}} & r_{int,sl}\sigma_{a_{int}}\sigma_{a_{sl}} & r_{int,d}\sigma_{a_{int}}\sigma_{a_d} \\ & \sigma^2_{a_{sl}} & r_{sl,d}\sigma_{a_{sl}}\sigma_{a_d} \\ Symmetric & & \sigma^2_{a_d} \end{bmatrix}$. $\sigma^2_{a_{int}}$, $\sigma^2_{a_{sl}}$ and $\sigma^2_{a_d}$ are the additive genetic variance of the intercept, slope and dispersion, respectively. $r_{int,sl}$, $r_{int,d}$ and $r_{sl,d}$ are the genetic correlations between the breeding values for intercept and slope of the reaction norm, intercept and dispersion, and slope and dispersion, respectively. The simulated macro-environmental effect was used as environmental covariate (**EC**) of the reaction norm and drawn from $N(0,\sigma^2_x)$. The pedigree used to construct **A** included generation 0 (base) and generation 1. Generation 0 consisted of unrelated sires for whom breeding values but not phenotypes were simulated. Generation 1 consisted of paternal half-sib groups of the base sires' offspring for whom both breeding values and phenotypes were simulated. Dams were not simulated.

Simulated data size and structure

Three different data structures were simulated: even, uneven and unbalanced.

The even data structure represented the ideal situation, where sires had equal number of offspring, macro-environments contained equal numbers of animals, and the sires' offspring were evenly distributed across macro-environments. The primary purpose of the even data structure was to serve as a comparison for the uneven and unbalanced data structures. Additionally, it was used to investigate the effects of progeny group size and total data size on estimation of variance components and EBVs.

The purpose of the uneven data structure was to investigate the effects of uneven use of sires on the estimation of variance components and EBVs. Hence, the uneven data structure differed from the even data structure in that sires had variable number of offspring. Fig. 1 shows the frequency of sires with between 1 and 200 offspring each in a pedigree provided by Angus Australia, which shows that most sires had few offspring and few sires had many offspring. To emulate this distribution, the number of offspring in each progeny group followed a gamma distribution

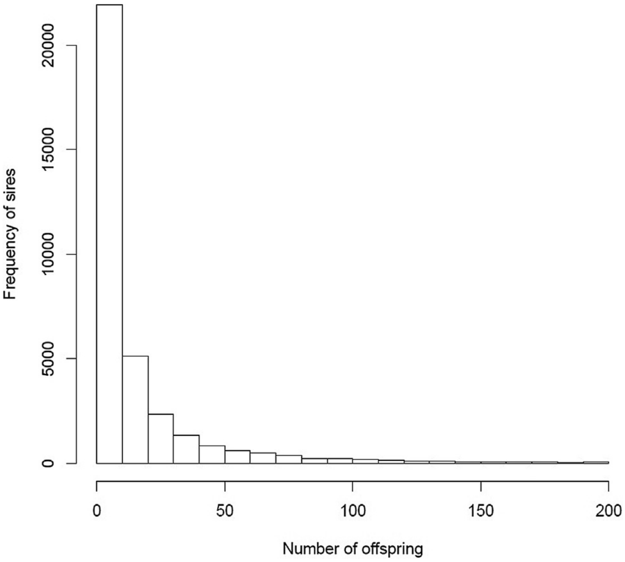

Fig. 1. The frequency of sires with 1–200 offspring in a Black Angus pedigree provided by Angus Australia.

$$\Gamma\left(1, n_{off}^{-1}\right) \tag{2}$$

where n_{off} was the average number of offspring per sire.

In order to examine the effects of unbalanced macro-environments, the unbalanced data structure had the same distribution of offspring per sire as the uneven data structure, but also had variable numbers of animals within macro-environments and the sires' offspring were unevenly distributed across macro-environments. The number of macro-environments a sire was to be represented in was drawn from

$$\Gamma\left(2, n_{macro}^{-2}\right) \tag{3}$$

where n_{macro} was the average number of animals within macro-environments. This distribution was chosen to ensure that most macro-environments contained few animals, similarly to the distribution of offspring per sire. However, while very large herds are few in numbers in most breeding programmes so are the very small herds, therefore, the most common number of animals within macro-environments was increased, compared to the most common number of offspring per sire, by using a shape parameter of 2 in Eq. (3) rather than 1 as in Eq. (2). The sires' offspring were randomly distributed over the number of macro-environments. The distribution assumptions resulted in a majority of small macro-environments and few large macro-environments. Thus, most sires had few offspring and were represented in few macro-environments and few sires had many offspring distributed over many macro-environments.

Within each simulated data structure, the total data size and progeny group size were varied to test the effects of progeny group sizes and the total data size alone and in combination with differences in data structure. The pedigree contained either 50, 100, 200, 500, or 1 000 sires with 20, 50, 100 or 200 offspring each (the even data structure) or on average (the uneven and unbalanced data structures).

To examine the impact of the genetic correlations ($r_{int,sl}$,$r_{int,d}$ and $r_{sl,d}$) deviating from 0 on the ability to estimate variance components, scenarios where the genetic correlations were 0.5 for either one or for all three genetic correlations were simulated for all three data structures with either 100 or 200 sires with either 100 or 200 offspring.

The parameters used in the simulation and their values are listed in Table 1. Values were chosen based on the values used by Mulder et al. (2013), but compared to their study, the number of sires was increased to examine the possibility of using datasets with more sires but fewer offspring per sire.

Statistical analysis

For all simulations, the data were analysed using the RN-DHGLM developed by Mulder et al. (2013). The RN-DHGLM was fitted at the level of sires rather than individual animals. The model was as follows:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y_d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu_d} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_{int}} & \boldsymbol{Z_{sl}} & 0 \\ 0 & 0 & \boldsymbol{Z_d} \end{bmatrix} \begin{bmatrix} \boldsymbol{s_{int}} \\ \boldsymbol{s_{sl}} \\ \boldsymbol{s_d} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_s} \\ \boldsymbol{e_{s_d}} \end{bmatrix} \tag{4}$$

where $\boldsymbol{y}$ and $\boldsymbol{y_d}$ were vectors containing response variables for the mean and dispersion parts (see below), respectively. $\boldsymbol{\mu}$ ($\boldsymbol{\mu_d}$) was the population mean for $\boldsymbol{y}$ ($\boldsymbol{y_d}$). $\boldsymbol{s_{int}}$ and $\boldsymbol{s_{sl}}$ were vectors of the additive genetic sire effects of the intercept and slope, respectively, of the reaction norm for $\boldsymbol{y}$ and $\boldsymbol{s_d}$ was a vector of the additive genetic sire effects for $\boldsymbol{y_d}$. $\boldsymbol{e_s}$ ($\boldsymbol{e_{s_d}}$) was a vector of the residuals for $\boldsymbol{y}$ ($\boldsymbol{y_d}$). $\boldsymbol{Z_{int}}$ ($\boldsymbol{Z_d}$) was a design matrix linking the response variable $\boldsymbol{y}$ ($\boldsymbol{y_d}$) to $\boldsymbol{s_{int}}$ ($\boldsymbol{s_d}$), and $\boldsymbol{Z_{sl}}$ was a column vector of the simulated macro-environmental effects (ECs). The distribution assumption for the random effects were $\begin{bmatrix} \boldsymbol{s_{int}} \\ \boldsymbol{s_{sl}} \\ \boldsymbol{s_d} \end{bmatrix} MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{s_{int}} & \sigma_{s_{int},s_{sl}} & \sigma_{s_{int},s_d} \\ \sigma_{s_{sl},s_{int}} & \sigma^2_{s_{sl}} & \sigma_{s_{sl},s_d} \\ \sigma_{s_d,s_{int}} & \sigma_{s_d,s_{sl}} & \sigma^2_{s_d} \end{bmatrix} \otimes \mathbf{A}\right)$ for additive genetic sire effects, where $\mathbf{A}$ is the additive genetic relationship matrix among sires, $\otimes$ was a Kronecker product, and $\begin{bmatrix} \boldsymbol{e_s} \\ \boldsymbol{e_{s_d}} \end{bmatrix} MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{I}\begin{bmatrix} \boldsymbol{W_s}^{-1} & 0 \\ 0 & \boldsymbol{W_{s_d}}^{-1} \end{bmatrix}\right)$ for residuals, where $\mathbf{I}$ was an identity matrix,$\boldsymbol{W_s} = diag\left(\widehat{\boldsymbol{y_d}}\right)^{-1}$ and $\boldsymbol{W_{s_d}} = diag\left(\frac{1-\boldsymbol{h}}{2}\right)$. $\boldsymbol{h}$ was the diagonal element of the part of the hat-matrix corresponding to $\boldsymbol{y}$. The hat-matrix ($\boldsymbol{H}$) is the matrix of leverages or information each phenotype provides to the predicted phenotype $\left(\begin{bmatrix} \widehat{\boldsymbol{y}} \\ \widehat{\boldsymbol{y_d}} \end{bmatrix} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y_d} \end{bmatrix}\right)$ (Mrode, 2005).

The response variables in $\boldsymbol{y}$ were the simulated phenotypes, while $\boldsymbol{y_d}$ contained linearised values of transformed squared residuals of $\boldsymbol{y}$

Table 1
Simulation parameters and their values.

Parameter	Value(s)
n_{sire}	50, 100, 200, 500, 1 000
n_{off}	20, 50, 100, 200
n_{macro}	100
$\sigma^2_{a_{int}}$	0.3
$\sigma^2_{a_{sl}}$	0.05
$\sigma^2_{a_d}$	0.1
$r_{int,sl}$, $r_{int,d}$,$r_{sl,d}$	0, 0.5
σ^2_x	1
$\sigma^2_{e_d}$	1
n_{rep}	100

n_{sire} = number of sires; n_{off} = average number of offspring per sire; n_{macro} = average number of animals in macro-environments; $\sigma^2_{a_{int}}$, $\sigma^2_{a_{sl}}$and$\sigma^2_{a_d}$ = the additive genetic variances of the intercept, slope, and dispersion, respectively; $r_{int,sl}$, $r_{int,d}$ and$r_{sl,d}$ = genetic correlations between the intercept and slope, intercept and dispersion, and slope and dispersion, respectively; σ^2_x = macro-environmental variance; $\sigma^2_{e_d}$ = environmental variance of the exponential model; n_{rep} = number of replicates.

$$y_{di} = \log\left(\bar{\sigma}^2_{e_s}\right) + \bar{\sigma}^{2\ -1}_{e_s} * \left(\frac{\hat{e}^2_i}{1 - h_i} - \bar{\sigma}^2_{e_s}\right) \quad (5)$$

where $\bar{\sigma}^2_{e_s} = \bar{w}_s{}^{-1} = \frac{tr(W_s)}{n}$ (n is the total number of records), $\hat{e}^2_i$ was the squared estimated residual of observation *i* and h_i was the leverage of observation *i* (Felleki et al., 2012).

All datasets were analysed using ASReml 4.1 (Gilmour et al., 2015).

The algorithm as implemented in ASReml 4.1 was an iterative weighted least squares approximation to the maximisation of the h-likelihood of the model (Rönnegård et al., 2010).

The algorithm was:

1. Initialise model by running a univariate reaction norm model on y, with homogeneous residual variance.
2. Calculate $\boldsymbol{y_d}$ and $\boldsymbol{W_{s_d}}$ and set $\boldsymbol{W_s} = diag\left(\hat{\sigma}^2_e\right)$
3. Run bivariate model in [4]
4. Update $\boldsymbol{W_s}$
5. Rerun the bivariate model
6. Update $\boldsymbol{y_d}$ and$\boldsymbol{W_{s_d}}$
7. Iterate steps 3–6 until converged.

See Supplementary Material S1 for an example of the ASReml scripts used to run the above algorithm.

Correction for use of sire model

By using a sire model, the estimated variance is based on sires (σ^2_s) and is thus only ¼ of the additive genetic variance (σ^2_a) (Mrode, 2005). The remaining ¾σ^2_a is included in the residual. Therefore, the estimated variances must be corrected for the additive genetic variance contained in the residual. For this, $\sigma^2_{s_{int}}$ and $\sigma^2_{s_{sl}}$ were multiplied by 4 to calculate $\sigma^2_{a_{int}}$ and $\sigma^2_{a_{sl}}$, respectively. However, because $\sigma^2_{a_d}$ was the genetic variance for the residual, and the estimated residual contained ¾$\sigma^2_{a_{int}}$ and ¾$\sigma^2_{a_{sl}}$, other corrections were necessary to obtain $\sigma^2_{a_d}$. In appendix 1 of Mulder et al. (2013), an algorithm for the sire model version of a RN-DHGLM was described. This algorithm included a correction for the fact that the residual variance of the mean part of the model contained ¾ of the additive genetic variance. In the suggested algorithm $\boldsymbol{y_d}$, $\boldsymbol{W_s}$ and $\boldsymbol{W_{s_d}}$ were calculated as:

$$y_{di} = \log\left(\bar{\sigma}^2_{e_s}\right) + \bar{\sigma}^{2\ -1}_{e_s} * \left(\frac{\hat{e}^2_i}{1 - h_i} * \frac{\bar{\sigma}^2_{e_s}}{\bar{\sigma}^2_{e_a}} - \bar{\sigma}^2_{e_s}\right) \quad (6)$$

$$\boldsymbol{W_s} = diag\left(\exp\left(log\left(\bar{\sigma}^2_{e_a}\right) + \boldsymbol{s_d}\right) + \frac{3}{4} * \left(\sigma^2_{a_{int}} + \sigma^2_{a_{sl}}\right)\right)^{-1} \quad (7)$$

$$\boldsymbol{W_{s_d}} = diag\left(\frac{(1 - \boldsymbol{h})^2}{2} * \left(\frac{\bar{\sigma}^2_{e_a}}{\bar{\sigma}^2_{e_s}}\right)^2\right) \quad (8)$$

where $\bar{\sigma}^2_{e_a} = \bar{\sigma}^2_{e_s} - \frac{3}{4} * \left(\sigma^2_{a_{int}} + \sigma^2_{a_{sl}}\right)$. These corrections result in an upwards scaling of the additive genetic variance of the dispersion. However, this algorithm is not implemented directly in ASReml4.1. Therefore, the following postestimation correction was used to obtain $\sigma^2_{a_d}$:

$$\sigma^2_{a_d} = \left(\frac{\bar{\sigma}^2_{e_a}}{\bar{\sigma}^2_{e_s}}\right)^2 * 4 * \sigma^2_{s_d} \quad (9)$$

Furthermore, treating the offspring's phenotypes as the phenotype of their sire, as done in this study, results in EBVs that are half of the true breeding values (**TBV**s). The EBVs for intercept (s_{int}) and slope (s_{sl})were therefore multiplied by 2. The EBVs of the dispersion were obtained using:

$$a_d = \frac{\sigma^2_{e_s}}{\sigma^2_{e_a}} * 2 * s_{sd} \quad (10)$$

Accuracy and bias of variance components and estimated breeding values

Within each replicate of each simulation, the accuracy and dispersion of the EBVs were assessed. The accuracy of the EBVs was evaluated as the correlation between the sires' TBVs and their EBVs, while the slope of a linear regression of sires' TBVs on their EBVs was considered the level of bias of the EBVs. A regression coefficient of <1 indicated overdispersion of the EBVs while underdispersion occurred when the regression coefficient was >1. Regression coefficients with 95% confidence intervals including 1 were considered unbiased.

For each scenario, the number of replicates for which the RN-DHGLM converged was calculated, along with the mean and empirical SEM of the estimated variance components ($\sigma^2_{a_{int}}$, $\sigma^2_{a_{sl}}$, $\sigma^2_{a_d}$, $r_{int,sl}$, $r_{int,d}$ and $r_{sl,d}$) and the mean and empirical SEM of the accuracy and dispersion of EBVs across all converged replicates.

Mean and empirical SEMs were calculated following:

$$mean = \frac{\sum_{l=1}^{n} value_l}{n} \quad (11)$$

$$SEM = \frac{SD}{\sqrt{n}} \quad (12)$$

where *value*$_l$ was the estimated variance component, prediction ability or bias of replicate *l*, *SD* was the SD of the mean $\left(SD = \sqrt{\sum_{l=1}^{n} \frac{|value_l - mean|^2}{n}}\right)$ and *n* was the number of converged replicates.

Results

Convergence

The percentage of replicates where the RN-DHGLM converged is presented in Fig. 2. The percentage of converged replicates increased with increasing number of sires and increasing number of offspring per sire, and the impact of the data structure decreased with increased data size. For 1 000 sires, all replicates converged, regardless of the number of offspring or the data structure. A high percentage of converged replicates, ranging from 87 to 100%, were obtained when the data had at least 100 sires with progeny group size of 50 or more, regardless of data structure. The percentage of converged replicates was not consistently higher for one data structure compared to the others. To summarise, the percentage of converged replicates did not seem to be consistently impacted by data structure, but small datasets had consistently lower percentage of converged replicates

Additive genetic variances

The additive genetic variance of intercept was accurately estimated across all data sizes and structures, with SEMs decreasing with increasing data size (Supplementary Table S1). For data with 20 offspring per sire, the additive genetic variance of slope and dispersion were consistently overestimated, when the number of sires

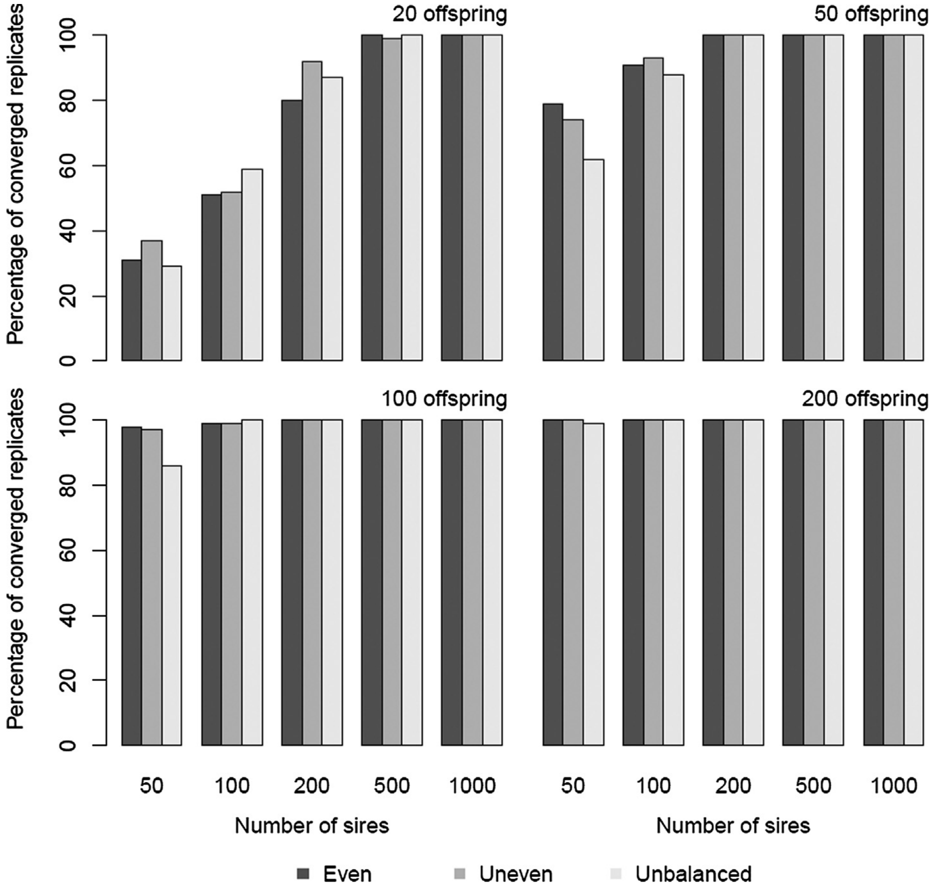

Fig. 2. Percentage of converged replicates out of 100 replicates for data with 50, 100, 200, 500 or 1 000 sires with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles).

was 200 or less (Figs. 3 and 4). With more offspring per sire and with more sires, the estimates became increasingly closer to the true value and the SEMs decreased. The impact of the data structures was not consistent across the number of sires or offspring. Due to the large range in the number of sires, results from datasets with 500 or 1 000 sires were not included in the figures, but can be found in Supplementary Table S1. The general trends observed when increasing the number of sires from 50 to 200 extend to datasets with 500 and 1 000 sires. These results showed that the data size had larger impact on the accuracy of variance components estimated using the RN-DHGLM than the data structure.

Genetic correlations

The estimated genetic correlations between intercept and slope, intercept and dispersion, and slope and dispersion (Supplementary Table S1) were generally not significantly different from the true values based on their 95% confidence intervals, regardless of the data size or structure. The SEMs decreased with an increasing number of sires or number of offspring per sire.

Increasing the true value of genetic correlations to 0.5, either one at a time or all three at once, for datasets with 100 or 200 sires and 100 or 200 offspring (Supplementary Table S2) resulted in similar accuracy and bias of variance components as observed with genetic correlations of 0. However, the percentage of converged replicates decreased from 99 to 100% when all genetic correlations were 0 to as low as 93% for the data with 100 sires and 100 offspring when the genetic correlation between slope and dispersion was 0.5, regardless of the size of the other genetic correlations.

Accuracy and bias of estimated breeding values

Increasing the number of offspring increased the accuracy of the EBVs (correlation between the TBVs and the EBVs of sires) and decreased the SEMs of the EBV accuracies for all three data structures (Figs. 5 and 6, Supplementary Table S3). The even data structure generally resulted in higher accuracy of the EBVs for intercept than the uneven or unbalanced data structures. Higher accuracies for the EBVs of the slope were also obtained from the even data structure for datasets with 50 offspring from 100 or more sires, or with 100 or more offspring regardless of the number of sires, compared to the uneven and unbalanced data structures with similar data sizes. The even data structure similarly resulted in higher accuracies for the EBVS of the dispersion, when data contained 50 offspring from 200, or more sires or at least 100 offspring regardless of the number of sires. The accuracies for the EBVs of the dispersion generally had overlapping 95% confidence intervals between the uneven and unbalanced data structures and neither of the data structures had consistently higher accuracy than the other across the varied data sizes. The trends in EBV accuracies observed when increasing the number of sires from 50 to 200 continue for 500 and 1 000 sires (Supplementary Table S3). The average number of offspring per sire had the largest impact on EBV accuracy; however, deviations from the ideal data structure (even data) resulted in reduced accuracies.

The regression of the TBVs on the EBVs (Supplementary Table S3) generally approached 1 (unbiased) with increasing data size, even though the EBVs for slope remained overdispersed (mean < 1 and confidence interval not containing 1) in most cases (Fig. 7). The trends for the regression coefficient of the TBVs on the

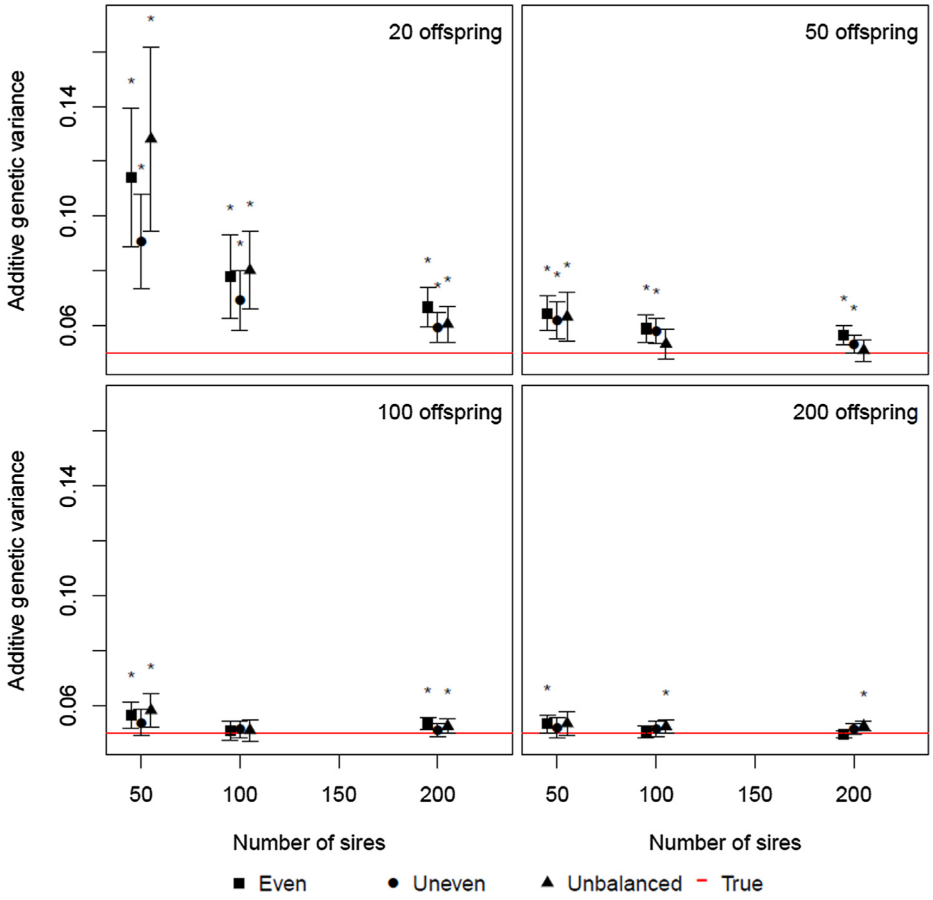

Fig. 3. Estimated additive genetic variance of slope for data with 50, 100 or 200 sires, with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles). Error bars indicate 95% confidence intervals and * indicates values where the confidence interval does not contain the true value (red line) of 0.05.

EBV continued when the datasets included 500 or 1 000 sires (Supplementary Table S3). There was no clear difference in the dispersion of the EBVs between the three data structures, indicating that data size, rather than data structure, affects the dispersion of the EBVs when using the RN-DHGLM.

The same patterns for EBV accuracies and bias were observed when the genetic correlations were increased, either individually or all three, to 0.5 for datasets with 100 or 200 sires and 100 or 200 offspring (Supplementary Table S4).

Discussion

Impacts of data size and structure

This study has illustrated that accuracy and dispersion of the EBVs and the estimation of variance components obtained using the RN-DHGLM are influenced by the size of the data. Small family sizes (20 or less offspring per sire) lead to overestimation of slope and dispersion, when data contained 200 or less sire families. The lack of enough data points to accurately estimate the slope and dispersion may have combined with the fact that small variances are more likely to be overestimated than underestimated, due to the lower boundary of 0 for variances. The study also showed that the data structure mostly affects EBV accuracy, while having little impact on the dispersion of the EBVs or the estimation of variance components. It has previously been shown that the precision of estimating variance components is lower for small datasets when using the RN-DHGLM (Mulder et al., 2013). As a result, Mulder et al. (2013) suggested that a minimum of 100 sires with at least 100 progeny would be required to estimate micro-GES with sufficient precision, measured as SD across replicates. However, the current study has shown that all variance components can be reasonably accurately estimated (means deviating ≤20% from true and SEMs ≤10% of true) from data with 20 offspring from 500 or more sires or 50 offspring from at least 100 sires. The current study used SEM rather than SD across replicates to evaluate the accuracy of the estimates. Using SEM takes the number of converged replicates into consideration and is thus a measure of the precision of the estimates, whereas SD across replicates is a measure of the variation in estimates. The difference in accuracy definitions may contribute to the difference in conclusions. Furthermore, Mulder et al. (2013) found that for datasets with 100 sires having 50 offspring or less each or datasets with 50 sires with 100 offspring, the algorithm had to force the covariance matrix to be positive definite for 5–51% of the replicates, respectively. In the current study, the covariance matrix was not forced to be positive definite, but the percentage of converged replicates followed a similar pattern to the number of replicates where the covariance matrix was bent to be positive definite in Mulder et al. (2013).

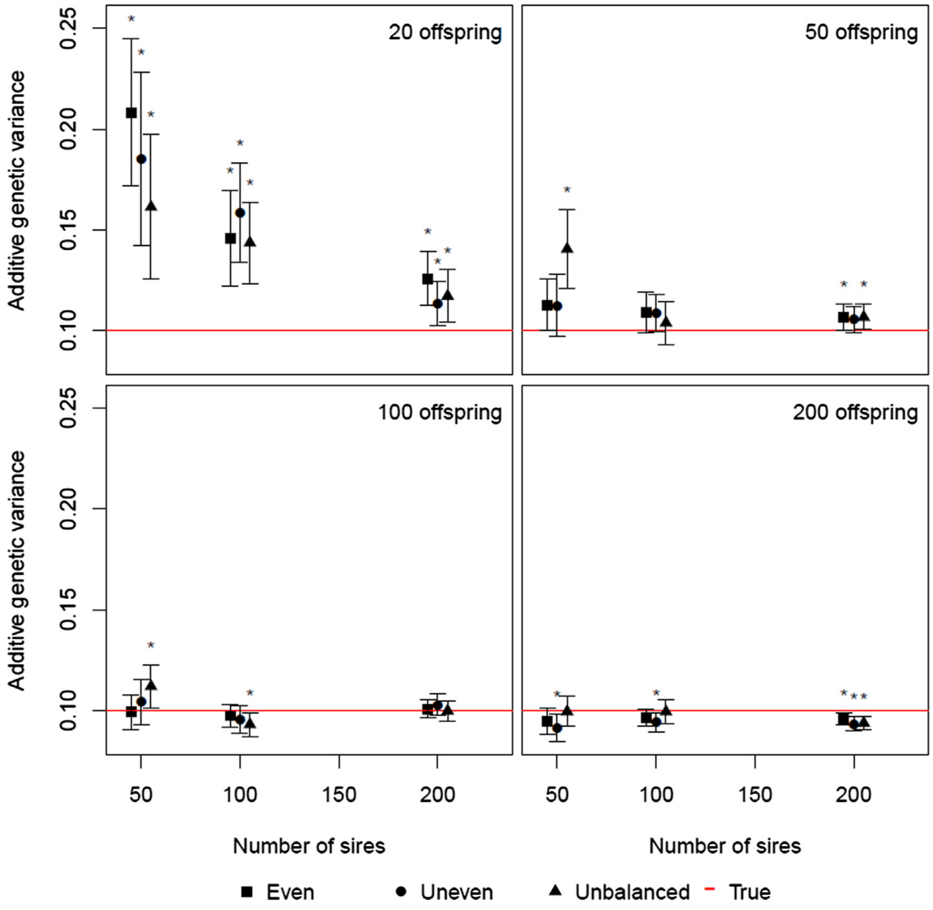

Fig. 4. Estimated additive genetic variance of dispersion for data with 50, 100 or 200 sires, with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles). Error bars indicate 95% confidence intervals and * indicates values where the confidence interval does not contain the true value (red line) of 0.1.

Small data sizes resulted in underdispersed EBVs for intercept and overdispersed EBVs for slope and dispersion, while the EBVs from larger datasets generally were unbiased (95% confidence interval overlapping 1) for intercept and dispersion. However, while the estimate was relatively close to 1 and stayed consistent across larger data sizes, the EBVs for slope of the reaction norm (macro-GES) were slightly overdispersed and had confidence intervals not overlapping 1 for some of the larger data sizes, because the increase in data resulted in reduced SEMs.

The accuracy of the EBV for each of the components (intercept, slope and dispersion) increased with the number of offspring per sire. This result aligns well with the expectation regarding the accuracy of a progeny test and the effective number of progeny per sire (Oldenbroek and Waaij, 2015). This also helps to explain why the even data structure generally had higher EBV accuracies than the uneven or unbalanced data. The accuracy of the EBVs is averaged across sires, so when the use of sires follows $\Gamma(1,n_{off}^{-1})$, a few sires will have a slight increase in accuracy due to having more offspring than the average. However, the increase will not be enough to offset the large number of sires with decreased accuracies due to having fewer offspring. Thus, the average accuracy is decreased in the datasets with uneven use of sires. The effect of uneven use of sires on the average accuracies of the EBVs is therefore as expected from the general linear mixed model theory.

Environmental covariate in reaction norms

Determining the EC to be used in a reaction norm is important when estimating parameters relating to macro-GES. In field data, the EC has to be estimated from the data itself. There are multiple examples where this was applied. Kolmodin et al. (2002) used deviations from the across-country average of the trait in question as primary herd environments for protein production and days open in Swedish dairy cattle. Fennewald et al. (2017) used within-region phenotypic averages as EC for birth and weaning weight. Li and Hermesch (2016) used the least squared means of herd by birth month. However, the use of pre-estimated or pre-calculated ECs results in an inclusion of a function of the data in the analysis which can result in biased estimates of variance components and EBVs. Alternatively, the model can iteratively update the EC throughout the analysis thereby adjusting for the other effects in the model (Calus et al., 2004; Su et al., 2006). Calus et al. (2004) found no increase in accuracy compared to using phenotypic herd-average, while Su et al. (2006) found the iterative procedure to be superior to the phenotypic mean. The procedure suggested by Su et al. (2006) was implemented in a Bayesian framework using Markov Chain Monte Carlo with Gibbs sampling, which means the error associated with estimated effects was integrated out before including it as EC. To reduce the risk of errors in

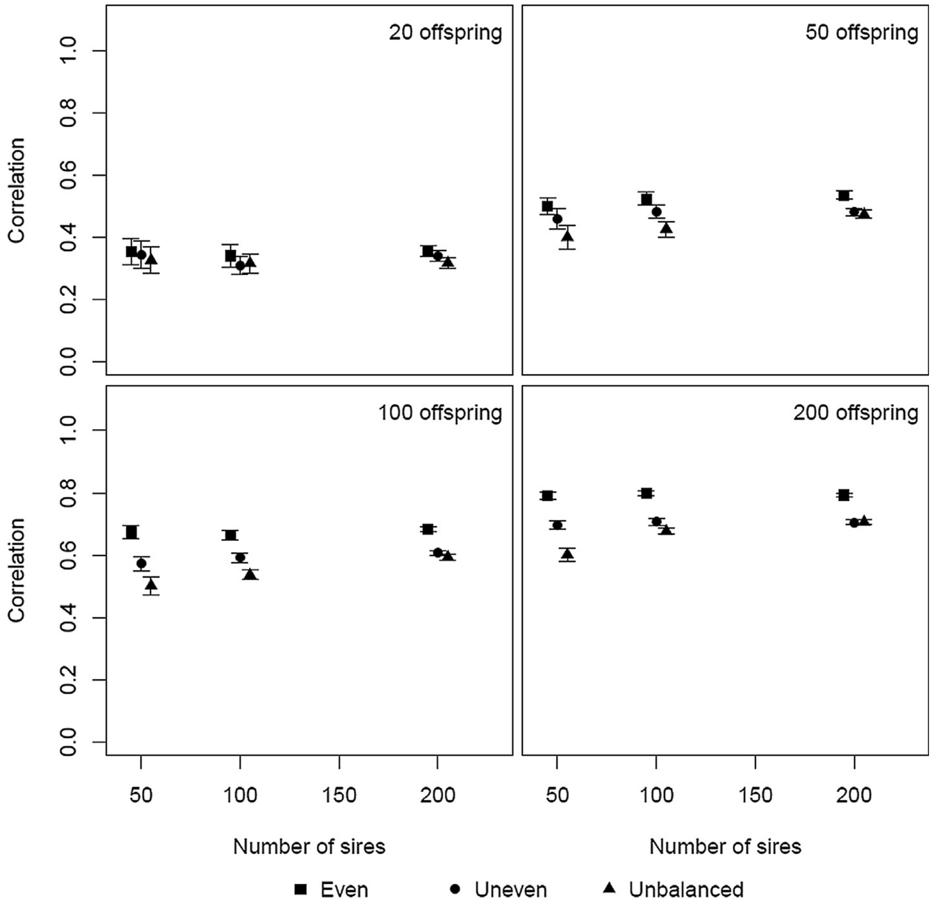

Fig. 5. Correlation between sires' true and estimated breeding values of slope for data with 50, 100 or 200 sires, with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles). Error bars indicate 95% confidence intervals.

the EC estimation influencing our comparisons, the simulated covariate was used, i.e. the macro-environmental effect similar to Mulder et al (2013).

However, in real data, EC values would have to be estimated and there could be some confounding between the genetic effects and the EC in unbalanced field data, it may be important to ensure that is minimised as this could influence the estimation.

Estimation with genetic correlations

For simplicity, the basic simulation used in this study assumed a zero correlation between each of the components; intercept, slope and dispersion. The genetic correlation was accurately estimated, regardless of the data structure, while the data size affected the SEMs, i.e. SEMs decreased with increasing data size. In real data, genetic correlations between intercept and slope have been found to range from 0.5 to 0.8 for milk yield in dairy cattle (Mulder et al., 2013; Ehsaninia et al., 2019), 0.3 for weaning weight and 0.7 for yearling weight in beef cattle (Bradford et al., 2016) and −0.1 to 0.7 for litter size in pigs (Knap and Su, 2008). It is important to note that the genetic correlation between intercept and slope depends on the placement of the intercept. The intercept is placed at 0 on the EC scale but depends therefore on whether and how the EC scale is scaled. In this study, the intercept was placed at the average EC which was 0. The genetic correlation between intercept and dispersion values has been found to range from −0.6 to −0.5 for litter size in pigs (Sorensen and Waagepetersen, 2003; Felleki et al., 2012), not significantly different from zero in fledgling weight in Great Tit (Mulder et al., 2016) and 0.5–0.7 for milk yield in dairy cattle (Mulder et al., 2013; Ehsaninia et al., 2019). The genetic correlations between slope and dispersion were 0.5–0.8 for milk yield in dairy cattle (Mulder et al., 2013; Ehsaninia et al., 2019). To examine the impact of non-zero genetic correlations, the simulated genetic correlations were raised to 0.5 between either intercept and slope, intercept and dispersion or slope and dispersion, or all three at once. When the correlation was increased to 0.5, there was no effect on the accuracy or bias of the estimated variance components or EBVs. In the current study, SEMs of the estimated genetic correlations did not increase compared to the scenarios with no genetic correlation. However, the percentage of converged replicates was lower when a genetic correlation of 0.5 was included between slope and dispersion, i.e. between macro- and micro-GES, regardless of whether the remaining genetic correlations were 0 or 0.5. Mulder et al. (2013) noted that the inclusion of a genetic correlation between the components can increase the number of replicates where the variance–covariance matrix needs to be forced to be positive definite in order to converge. Overall, the presence of genetic correlations of 0.5 had only a slight impact on the estimation of variance components or accuracy and bias of EBVs using the RN-DHGLM.

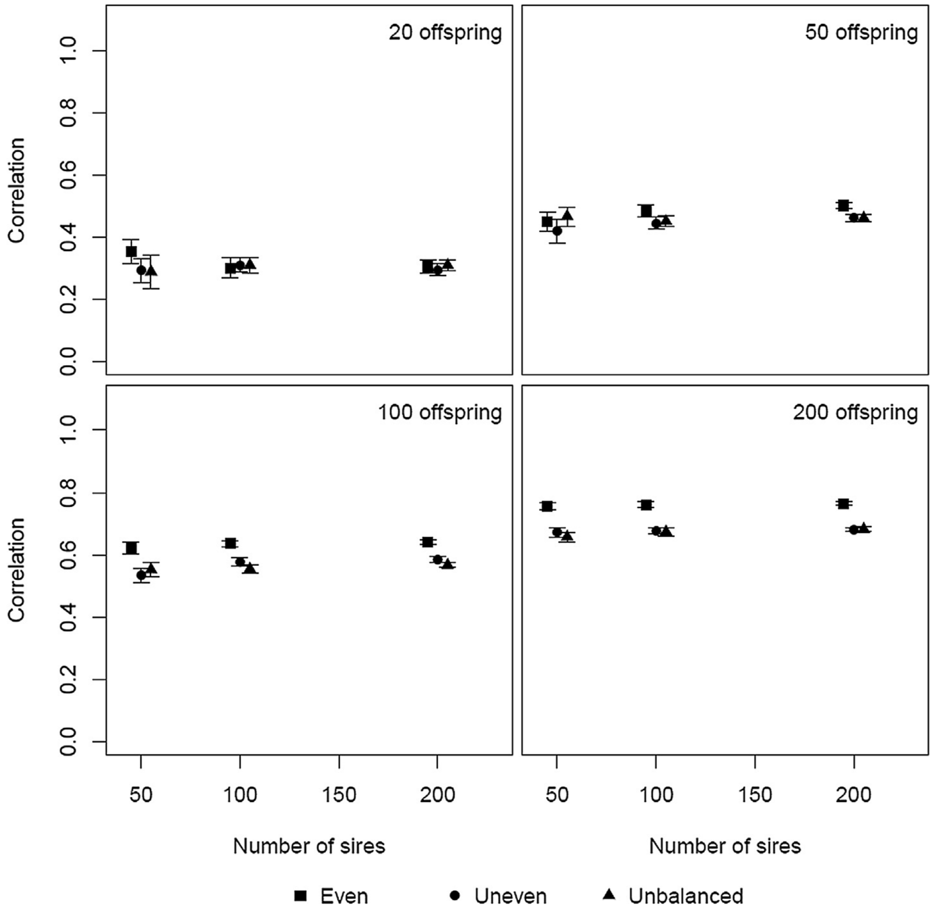

Fig. 6. Correlation between sires' true and estimated breeding values of dispersion for data with 50, 100 or 200 sires, with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles). Error bars indicate 95% confidence intervals.

Application to real data

This study has shown that macro- and micro-environmental sensitivity can be estimated when the dataset is more unbalanced than previously suggested. This finding allows for the methods to be applied in real data scenarios. Real datasets may enable the study of interesting concepts that are often difficult to simulate. An example of this is environmentally heterogeneous residual variance which was not included in the current study, due to the complexity it would add to both the simulations and analysis. In the current study, the residuals have genetically heterogeneous but environmentally homogeneous residual variance. Many studies have shown traits in multiple species with environmentally heterogeneous residual variance (e.g. Fujii and Suzuki, 2006; Shirali et al., 2015; Madsen et al., 2018), and therefore, the impacts of heterogeneous residual variance should be investigated in future studies. The current study also included a single generation of progeny, with a complete half-sib design (assuming sires are mated to random dams), and did not examine the effects of selection and non-random mating. Real datasets typically cover multiple, and often overlapping, generations and include selection and non-random mating. Interestingly, the RN-DHGLM is a special type of a mixed model that has the properties of BLUP, therefore, it is likely that multiple and overlapping generations, selection and non-random mating would be accurately handled provided the data and pedigree is appropriate and correctly recorded (Mrode, 2005). Furthermore, fixed environmental effects other than the population mean are expected in real data. Mulder et al. (2013) randomly assigned animals to contemporary groups within macro-environments without simulating the effect of these groups. When contemporary group was included as a fixed effect (data with 100 sire with 100 offspring each), an increase was observed for the number of replicates where the covariance matrix had to be forced to be positive definite as well as the SD of the estimated genetic variances. The estimated variance themselves changed little compared to those from analyses without the fixed effect of contemporary group (Mulder et al., 2013). The need to force the covariance matrix of more replicates to be positive definite may stem from the fact that the fixed effect is absorbing part of the additive genetic effect, especially with small contemporary groups. Therefore, the estimated genetic variance may become lower resulting in a higher probability that the genetic variance–covari-

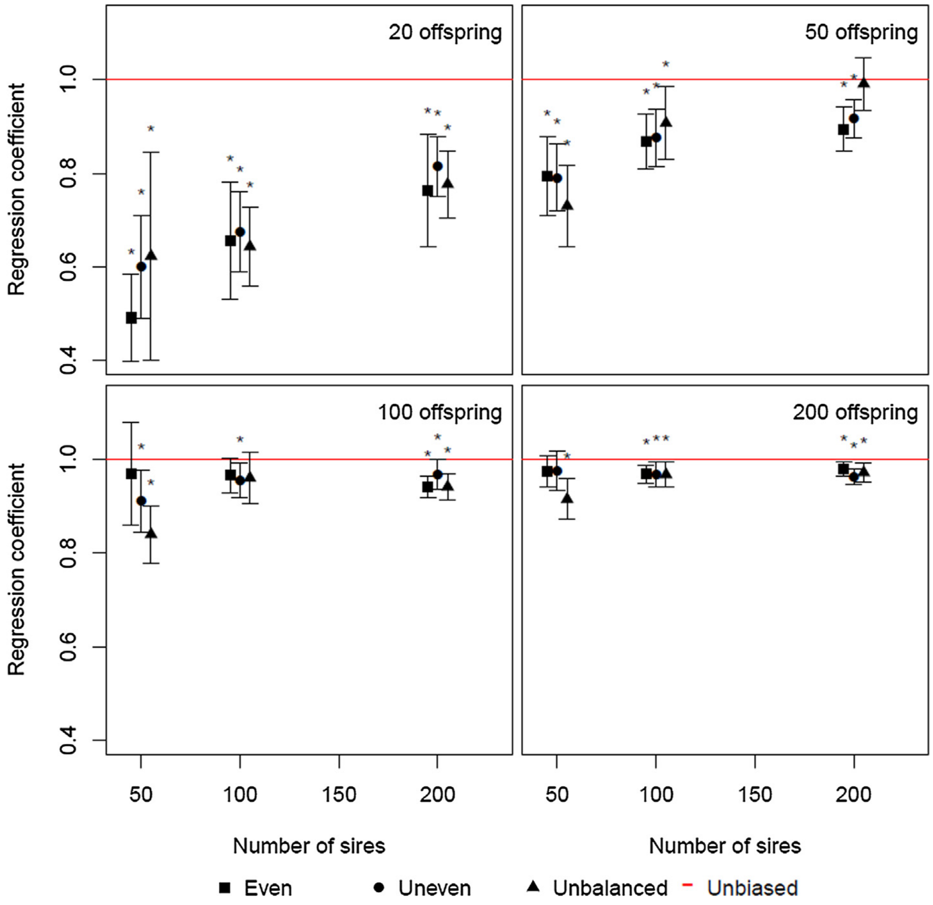

Fig. 7. Regression coefficient from a regression of sires' true breeding values on their estimated breeding values of slope for data with 50, 100 or 200 sires, with varied number of offspring per sire for the even data structure (squares), the uneven data structure (circles) and unbalanced data structure (triangles). Error bars indicate 95% confidence intervals and * indicates values where the confidence interval does not contain 1 (red line).

ance matrix is not positive definite. The findings from this study provide an opportunity to examine the estimation of macro- and micro-environmental sensitivity in real data that cover varied environments and production systems.

Conclusions

Both accuracy and bias of the EBVs and the variance component estimation were influenced by the data size, while the data structure only consistently affected accuracy of EBVs. The inclusion of a positive genetic correlation between genetic components did not influence the accuracy and bias of the EBVs or the estimation of variance components, regardless of the data structure. The results show the RN-DHGLM is applicable to unbalanced data and data with small average family size (20 offspring on average) when the number of sire families is large (minimum 500) or data with larger sire family size (at least 50) and fewer sire families (100 or more). The possibility of using the RN-DHGLM on unbalanced data and on data with relatively small sire families allows for the estimation of macro- and micro-GES in larger field datasets and obtaining macro- and micro-GES EBVs for sires with few offspring. This increases the number of candidates that can be potentially selected to improve macro- and micro-GES within a breeding programme.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.animal.2021.100411.

Ethics approval

The data was simulated, therefore no ethical approval were necessary.

Data and model availability statement

None of the data were deposited in an official repository. Inquiries should be made to the corresponding author.

Author ORCIDs

MDM: https://orcid.org/0000-0003-4997-6779, **JVDW:** https://orcid.org/0000-0003-2512-1696, **VB:** https://orcid.org/0000-0001-8005-9253, **HAM:** https://orcid.org/0000-0003-2124-4787, **SC:** https://orcid.org/0000-0001-8605-1738.

Author contributions

MDM generated and analysed the data and is main author of the manuscript, **SC** aided with generation and analysis of data and was a major contributor in writing the manuscript. **SC, JVDW, VB** and **MDM** conceived and designed the simulations. **HAM** aided with data analysis. All authors read and approved the final manuscript.

Declaration of interest

The authors declare that they have no competing interests.

Acknowledgements

The authors thank A. Gilmour and P. Madsen for their contributions towards simulation and analyses.

Financial support statement

M. D. Madsen acknowledges the Livestock Genomics Scholarship provided by University of New England for funding their studies.

References

Bradford, H.L., Fragomeni, B.O., Bertrand, J.K., Lourenco, D.A.L., Misztal, I., 2016. Genetic evaluations for growth heat tolerance in Angus cattle. Journal of Animal Science 94, 4143–4150. https://doi.org/10.2527/jas.2016-0707.

Calus, M.P.L., Bijma, P., Veerkamp, R.F., 2004. Effects of data structure on the estimation of covariance functions to describe genotype by environment interactions in a reaction norm model. Genetics, Selection, Evolution 36, 489–507. https://doi.org/10.1186/1297-9686-36-5-489.

Ehsaninia, J., Ghavi Hossein-Zadeh, N., Shadparvar, A.A., 2019. Estimation of genetic variation for macro- and micro-environmental sensitivities of milk yield and composition in Holstein cows using double hierarchical generalized linear models. Journal of Dairy Research 86, 145–153. https://doi.org/10.1017/S0022029919000293.

Falconer, D.S., Mackay, T.F.C., 1996. Chapter 8 Variance. In: Falconer, D.S. (Ed.), Introduction to Quantitative Genetics. Pearson Education Limited, Essex, United Kingdom, pp. 125–147.

Felleki, M., Lee, D., Lee, Y., Gilmour, A.R., Rönnegård, L., 2012. Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models. Genetic Research 94, 307–317. https://doi.org/10.1017/S0016672312000766.

Fennewald, D.J., Weaber, R.L., Lamberson, W.R., 2017. Genotype by environment interactions for growth in Red Angus. Journal of Animal Science 95, 538–544. https://doi.org/10.2527/jas.2016.0846.

Fujii, C., Suzuki, M., 2006. Comparison of homogeneity and heterogeneity of residual variance using random regression test-day models for first lactation Japanese holstein cows. Animal science journal 77, 28–32. https://doi.org/10.1111/j.1740-0929.2006.00316.x.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J., Thompson, R., 2015. ASReml User Guide Release 4.1 Functional Specification. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Hill, W.G., Mulder, H.A., 2010. Genetic analysis of environmental variation. Genetic Research 92, 381–395. https://doi.org/10.1017/S0016672310000546.

Knap, P.W., Su, G., 2008. Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis. Animal 2, 1742–1747. https://doi.org/10.1017/S1751731108003145.

Kolmodin, R., Strandberg, E., Madsen, P., Jensen, J., Jorjani, H., 2002. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. Acta Agriculturae Scandinavica, Section A - Animal Science 52, 11–24. https://doi.org/10.1080/09064700252806380.

Lee, Y., Nelder, J.A., 2006. Double hierarchical generalized linear models (with discussion. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55, 139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x.

Li, L., Hermesch, S., 2016. Evaluation of sire by environment interactions for growth rate and backfat depth using reaction norm models in pigs. Journal of Animal Breeding and Genetics 133, 429–440. https://doi.org/10.1111/jbg.12207.

Madsen, M.D., Madsen, P., Nielsen, B., Kristensen, T.N., Jensen, J., Shirali, M., 2018. Macro-environmental sensitivity for growth rate in Danish Duroc pigs is under genetic control. Journal of Animal Science 96, 4967–4977. https://doi.org/10.1093/jas/sky376.

Mrode, R.A., 2005. Linear models for the prediction of animal breeding values. CAB International, Biston, MA, USA.

Mulder, H., Rönnegård, L., Fikse, W., Veerkamp, R., Strandberg, E., 2013. Estimation of genetic variance for macro- and micro-environmental sensitivity using double hierarchical generalized linear models. Genetics, Selection, Evolution 45, 23. https://doi.org/10.1186/1297-9686-45-23.

Mulder, H.A., Gienapp, P., Visser, M.E., 2016. Genetic variation in variability: Phenotypic variability of fledging weight and its evolution in a songbird population. Evolution 70, 2004–2016. https://doi.org/10.1111/evo.13008.

Oldenbroek, K., Waaij, L.V.D., 2015. Chapter 8.4 Accuracy of the breeding value; the basic concept. In Textbook Animal Breeding and Genetics for BSc students (ed. Kennisnet, G.). Centre for Genetic Resources, The Netherlands and Animal Breeding and Genomics Centre, Wageningen, The Netherlands, pp 167–168.

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H.A., Strandberg, E., 2010. Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. Genetics, Selection, Evolution: GSE 42, 8–8. https://doi.org/10.1186/1297-9686-42-8.

Rönnegård, L., Felleki, M., Fikse, W.F., Mulder, H.A., Strandberg, E., 2013. Variance component and breeding value estimation for genetic heterogeneity of residual variance in Swedish Holstein dairy cattle. Journal of Dairy Science 96, 2627. https://doi.org/10.3168/jds.2012-6198.

SanCristobal-Gaudy, M., Elsen, J.-M., Bodin, L., Chevalet, C., 1998. Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. Genetics, Selection, Evolution 30, 423–451. https://doi.org/10.1186/1297-9686-30-5-423.

Shirali, M., Nielsen, V.H., Møller, S.H., Jensen, J., 2015. Longitudinal analysis of residual feed intake and BW in mink using random regression with heterogeneous residual variance. Animal 9, 1597–1604. https://doi.org/10.1017/S1751731115000956.

Sorensen, D., Waagepetersen, R., 2003. Normal linear models with genetically structured residual variance heterogeneity: a case study. Genetic Research 82, 207–222. https://doi.org/10.1017/S0016672303006426.

Su, G., Madsen, P., Lund, M.S., Sorensen, D., Korsgaard, I.R., Jensen, J., 2006. Bayesian analysis of the linear reaction norm model with unknown covariates. Journal of Animal Science 84, 1651–1657. https://doi.org/10.2527/jas.2005-517.