CBCS
HTTP://AGROBIOL.SGGW.WAW.PL/CBCS

COMMUNICATIONS IN BIOMETRY AND CROP SCIENCE
VOL. 9, NO. 1, 2014, PP. 26–40

INTERNATIONAL JOURNAL OF THE FACULTY OF AGRICULTURE AND BIOLOGY,
WARSAW UNIVERSITY OF LIFE SCIENCES – SGGW, POLAND

REGULAR ARTICLE

Analysis of a complex trait with missing data on the component traits

Hans-Peter Piepho [1]*, Bettina U. Müller [2], Constantin Jansen [2]

[1] Bioistatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany.
[2]Strube Research GmbH & Co. KG, Hauptstraße 1, 38387 Söllingen, Germany.
*Corresponding author: Hans-Peter Piepho; E-mail: piepho@uni-hohenheim.de

CITATION: Piepho, H.P., Müller, B.U., Jansen, C. (2014). Analysis of a complex trait with missing data on the component traits. *Communications in Biometry and Crop Science* 9 (1), 26–40.

Received: 24 July 2014, Accepted: 2 October 2014, Published online: 28 October 2014

ABSTRACT

Many complex agronomic traits are computed as the product of component traits. For the complex trait to be assessed in a field plot, each of the component traits needs to be measured in the same plot. When data on one or several component traits are missing, the complex trait cannot be computed. If the analysis is to be performed on data for the complex trait, plots with missing data on at least one of the component traits are discarded, even though data may be available on some of the component traits. This paper considers a multivariate mixed model approach that allows making use of all available data. The key idea is to employ a logarithmic transformation of the data in order to convert a product into a sum of the component traits. The approach is illustrated using a series of sunflower breeding trials. It is demonstrated that the multivariate approach allows making use of all available information in the case of missing data, including plots that may have data only on one of the component traits.

Key Words: *logarithmic transformation; log-normal distribution; multiplicative model; multi-trait analysis; multivariate mixed model; yield component analysis; yield components.*

INTRODUCTION

Yield is a complex trait that can be expressed as the product of yield components (Fraser and Eaton 1983). For example, the complex trait 'dry oil yield' (DOY) in sunflower can be computed as the product of the yield components 'oil content' (OC) and 'dry matter yield' (DY). In the analysis of randomized field trials, DOY can be computed per plot and then subjected to univariate analysis of variance (ANOVA) by a suitable linear model.

Quite frequently, data are missing on some plots for one of the two component traits (OC, DY), meaning that DOY cannot be computed for these plots, leading to missing data in the analysis for DOY. Most notably, it is good practice to check the data for approximate

normality and discard any outliers, which will routinely lead to missing values for some traits on some plots. Hence, incomplete data for component traits can occur in practice. If DOY data per plot are to be analysed, the available information on plots with missing data on one of the component traits is discarded. This suggests that a more efficient analysis would be possible if all available data could be analysed, including data on such plots where information on one of the component traits is missing. One option is to perform a bivariate analysis of the component traits and then derive the analysis for the composite trait DOY from that bivariate analysis. The advantage of this approach is that all available data can be utilized, including data from plots that have observations on only one of the component traits. While this idea is very appealing and may seem straightforward at first sight, it involves some challenges, because it is not immediately obvious how a joint model for the complex trait (DOY) and its components should be formulated.

To illustrate the problem, consider an experiment laid out according to a randomized complete block design (RCBD). Analysis for each component traits c $(c = 1,2...,C)$ could be based on the model

$$x_{ij}^{(c)} = \mu^{(c)} + g_i^{(c)} + b_j^{(c)} + e_{ij}^{(c)} \ , \tag{1}$$

where $\mu^{(c)}$ is an intercept, $g_i^{(c)}$ is the effect of the *i*-th genotype, $b_j^{(c)}$ is the effect of the *j*-th block, and $e_{ij}^{(c)}$ is the *ij*-th plot error, which is usually assumed to have a normal distribution with constant variance. Under this model for the component traits (here: OC, $c = 1$; DY, $c = 2$), the model for the response y_{ij} of the complex trait (DOY) could be derived by multiplication as

$$y_{ij} = x_{ij}^{OC} \times x_{ij}^{DY} = \left(\mu^{OC} + g_i^{OC} + b_j^{OC} + e_{ij}^{OC}\right) \times \left(\mu^{DY} + g_i^{DY} + b_j^{DY} + e_{ij}^{DY}\right). \tag{2}$$

Resolving the right-hand side of (2) obviously leads to a model with several multiplicative terms, so the model implied for the complex trait DOY by the multiplicative relation in (2) is more complicated than model (1), which is assumed for the component traits OC and DY. For example, one of the emerging multiplicative terms is the product of the error terms of the two component traits $\left(e_{ij}^{OC} \times e_{ij}^{DY}\right)$. This product clearly has a non-normal distribution if the two component error terms have a normal distribution. Moreover, the product in (2) involves cross products $g_i^{OC} b_j^{DY}$ and $g_i^{DY} b_j^{OC}$, which correspond to an interaction of the block and treatment factors. But model (1), which is routinely used for analysis also of the complex trait DOY, assumes normality of errors and absence of block-treatment interaction. So there is obviously a conflict of model assumptions, if the same type of linear model (eq. 1) is to be used for both the complex trait as well as for each of its component traits.

In this paper we will show that a resolution of the conflict of model assumptions is forthcoming by simply conducting all analyses on the logarithmic scale (Piepho 1995, Gołaszewski 1996, Kozak 2004, Kozak and Mądry 2006). We discuss implications of this approach for the assumed distribution of traits on the original scale. The approach is illustrated using field trial data from a sunflower breeding programme.

MATERIALS AND METHODS

THE MODEL

The complex trait Y (DOY in our example) can be expressed as the product of its component traits $\left(X^{(1)}, X^{(2)},...\right)$, i.e.,

$$Y = X^{(1)} \times X^{(2)} \times .... \ . \tag{3}$$

By taking logarithms, the multiplicative model (3) is converted into an additive relation (Piepho 1995):

$$\log(Y) = \log\left(X^{(1)}\right) + \log\left(X^{(2)}\right) + \ldots , \tag{4}$$

where $\log(\ )$ denotes the natural logarithm (i.e., logarithm with base equal to e = Euler's constant; other bases can be used, but note that variance components will change by a common scaling factor). To keep the notation simple, we may express (4) as

$$\tilde{Y} = \tilde{X}^{(1)} + \tilde{X}^{(2)} + \ldots = \sum_{c=1}^{C} \tilde{X}^{(c)} , \tag{5}$$

where $\tilde{Y} = \log(Y)$ and $\tilde{X}^{(c)} = \log\left(X^{(c)}\right)$ $(c = 1,2,\ldots,C)$. We assume a linear model to hold on the logarithmic scale. For illustration, consider the oil yield example given in the introduction, were we consider an experiment laid out according to an RCBD. The model for the c-th component trait can be written as

$$\tilde{x}_{ij}^{(c)} = \mu^{(c)} + g_i^{(c)} + b_j^{(c)} + e_{ij}^{(c)} . \tag{6}$$

With this model for the component traits, the model for the complex trait (DOY) can be written in terms of sums (rather than products because we have moved to a logarithmic scale) as follows:

$$\tilde{y}_{ij} = \mu + g_i + b_j + e_{ij} , \tag{7}$$

where $\tilde{y}_{ij} = \sum_{c=1}^{C} \tilde{x}_{ij}^{(c)}$, $\mu = \sum_{c=1}^{C} \mu^{(c)}$, $g_i = \sum_{c=1}^{C} g_i^{(c)}$, $b_j = \sum_{c=1}^{C} b_j^{(c)}$, and $e_{ij} = \sum_{c=1}^{C} e_{ij}^{(c)}$. It is seen that the model for the complex trait is of the same form as the model for the component traits, so the conflict described in the introduction is nicely resolved. This important property suggests that we can fit a multivariate linear model to the component trait vector $\tilde{\mathbf{X}} = \left(\tilde{X}^{(1)}, \tilde{X}^{(2)}, \ldots\right)^T$ and then derive any inferences with respect to the complex trait from the multivariate model fitted to the component traits. In particular, it emerges that on the logarithmic scale the adjusted genotype mean for the complex trait can simply be computed as the sum of the corresponding adjusted genotype means for the component traits. The variance components of random effects can be similarly derived, as will be shown next.

Continuing with the RCBD example, the multivariate model can be written as

$$\tilde{\mathbf{x}}_{\mathbf{ij}} = \boldsymbol{\mu} + \mathbf{g}_{\mathbf{i}} + \mathbf{b}_{\mathbf{j}} + \mathbf{e}_{\mathbf{ij}} , \tag{8}$$

where $\tilde{\mathbf{x}}_{\mathbf{ij}} = \left(\tilde{x}_{ij}^{(1)}, \tilde{x}_{ij}^{(2)}, \ldots\right)^T$, $\boldsymbol{\mu} = \left(\mu^{(1)}, \mu^{(2)}, \ldots\right)^T$, $\mathbf{g}_{\mathbf{i}} = \left(g_i^{(1)}, g_i^{(2)}, \ldots\right)^T$, $\mathbf{b}_{\mathbf{j}} = \left(b_j^{(1)}, b_j^{(2)}, \ldots\right)^T$ and $\mathbf{e}_{\mathbf{ij}} = \left(e_{ij}^{(1)}, e_{ij}^{(2)}, \ldots\right)^T$. Now assume that genotype and error effects are multivariate normal, i.e., $\mathbf{g}_{\mathbf{i}} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{g}}\right)$ and $\mathbf{e}_{\mathbf{ij}} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{e}}\right)$. Then the genetic and error variances of the complex trait (DOY) are given by $\sigma_g^2 = \mathbf{1}^{\mathbf{T}} \boldsymbol{\Sigma}_{\mathbf{g}} \mathbf{1}$ and $\sigma_e^2 = \mathbf{1}^{\mathbf{T}} \boldsymbol{\Sigma}_{\mathbf{e}} \mathbf{1}$, respectively, where $\mathbf{1}$ is a c-dimensional column vector of ones. These ideas are readily extended to any other design and form of linear model. The key point is that on a logarithmic scale the complex trait is the simple sum of the component traits, so all inference for the complex trait as derived from the multivariate model for the component traits is essentially linear, which greatly simplifies the analysis compared to an analysis on the original scale.

For a REML-based analysis of unbalanced data with missing values, it must be assumed that the missing data mechanism meets the missing-at-random (MAR) assumption (Piepho

and Möhring 2006). In the following, we compare univariate analyses for the complex trait DOY with various multivariate analyses.

In summary, we are assuming an additive model for the yield components on the logarithmic scale. The response of complex trait is the sum of the responses of the component traits on the logarithmic scale. On the original scale, the model becomes multiplicative in the exponentiated effects from logarithmic scale. Thus, for emphasis of this property, we may simply refer to our model as the "multiplicative model".

IMPLEMENTATION OF THE MULTIPLICATIVE MODEL IN A MIXED MODEL PACKAGE

To exemplify implementation of the multiplicative model in a linear model package, for simplicity we here consider the RCBD example with two traits. Extension to more complex settings is straightforward. For multivariate analysis, we assume that the data are arranged in the format illustrated in Table 1, where for an observational unit (plot) each trait is represented by a separate record.

Table 1. Coding of variables§ for the first three plots in a dataset for experiment laid out as RCBD for two component traits oil content (OC; in %) and dry yield (DY; in t/ha).

Blk	Plt	Gen	Trait	Y (response)
1	1	8	OC	48.8
1	1	8	DY	2.1
1	2	3	OC	49.1
1	2	3	DY	1.9
1	3	17	OC	51.2
1	3	17	DY	1.8
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

§: Blk = block, Plt = plot ID, Gen = genotype, Trait = trait.

For a single trait, the model can be expressed as (Piepho et al. 2003)

$$\log(Y) = \text{Blk} : \text{Gen} + \underline{\text{Plt}} \ , \qquad (9)$$

where the fixed block effect (Blk) appears before the colon and the random effects for genotype (Gen) and error (Plt) after the colon. The residual error term is underscored. Note that the symbolic representation of the model in (9) is entirely equivalent to the more standard representation in (6), where each effect, except for the intercept, has specific subscripts (Piepho et al. 2003). We use the symbolic form for further development for simplicity because this avoids having to write down many subscripts and because it is closer to the syntax required for implementation in a mixed model package. Note that there is no loss of information in the transition form the "usual" form to the symbolic form.

The univariate model (9) may be extended to the multivariate case as

$$\log(Y) = \text{Blk.Trait} : \boldsymbol{Gen}\textbf{.Trait} + \underline{\boldsymbol{Plt}}\textbf{.Trait} \ . \qquad (10)$$

In (10), we have used a notation that was suggested by Piepho et al. (2004) in the context of repeated measures designs. The situation here is analogous with traits corresponding to repeated measures. Thus, we boldfaced and italicised the random effect that would be fitted to a single trait, whereas the "repeated" factor Trait is only boldfaced. The former part of the effect identifies the "subject" on which repeated measurements, or measurements on multiple traits, are taken (Verbeke and Molenberghs 1997). This model is easily implemented in a mixed model package (Piepho and Möhring 2011). For example, the SAS code for model (10), assuming an unstructured variance-covariance model for both the error and the

genotypic effect, is given in Box 1. We here use the UNR structure, which models the variance-covariance matrix using the variances of the traits and the pairwise correlations.

Box 1. SAS code for fitting the multivariate model (10).

```
proc mixed;
class blk plt gen trait;
model log_y=blk*trait;
random   trait / subject=gen type=unr;
repeated trait / subject=plt type=unr;
run;
```

In order to compute adjusted means for genotypes, the genotype effect may be taken as fixed. It is then also convenient to introduce main effects for both blocks and genotypes, because this facilitates the computation of genotype means across traits, from which the corresponding sums can be obtained in a simple post-processing step explained below:

log(Y) = Blk + Blk.Trait + Gen + Gen.Trait : ***Plt*.Trait** (11)

This model may be fitted in SAS using the directives in Box 2. The adjusted mean of the genotypes for the complex trait is equal to the sum of the corresponding means for the component traits. This sum is equal to the genotype mean across traits under model (11), which is easily computed using the code in Box 2, multiplied by the number of traits.

Box 2. SAS code for fitting the multivariate model (11) and computing genotype means across traits.

```
proc mixed;
class blk plt gen trait;
model log_y=blk blk*trait gen gen*trait;
repeated trait / subject=plt type=unr;
lsmeans gen;
run;
```

BACK TO THE ORIGINAL SCALE

For selection, means on the transformed logarithmic scale are sufficient, because back-transformation to the original scale would not affect genotype ranking. For interpretation, however, one may be interested in mean estimates on the original scale. If means on the logarithmic scale are naïvely back-transformed by applying the exponential function, we obtain estimates of medians (Piepho 2009) under the multiplicative model, but not of the expected values on the original scale. For estimating the expected values, we may make use of properties of the log-normal distribution (Johnson et al., 1994). If μ and σ^2 are the expected value and variance, respectively, of the response $\log(y)$, then the expected value on the original scale is $E(y) = \exp(\mu + \sigma^2/2)$, whereas the median equals $Median(y) = \exp(\mu)$. To estimate the expected value on the original scale, we may simply plug in the adjusted genotype mean on the logarithmic scale for μ and set σ^2 equal to the total variance of an observation on the logarithmic scale. Assuming a mixed model with simple random effects and constant variances, the total variance will be a constant for all observations. Note that $E(y) = \exp(\sigma^2/2) \times Median(y)$, meaning that for a constant total variance σ^2, and hence a

constant factor of proportionality $\exp(\sigma^2/2)$, the expected value and the median are perfectly correlated. It emerges that, as far as the correlation and ranking of genotypes are concerned, no additional information is forthcoming by estimating the expected value rather than the median on the original scale. Thus, all analyses can be performed on the logarithmic scale. If a back-transformation is needed for ease of interpretation and descriptive purposes, it is sufficient to naïvely back-transform adjusted means for genotypes to yield median estimates on the original scale.

APPROXIMATE NORMALITY ON THE ORIGINAL SCALE?

It may be of interest to assess the degree to which the assumption of approximate normality on the original scale (that is usually made without any further ado) is commensurate with the assumption of normality on the logarithmic scale. Assuming normality on the logarithmic scale implies log-normality on the original scale. But when the variance on the logarithmic scale (σ^2) is small, the log-normal distribution on the original scale is close to normal. Specifically, the skewness is $\left[\exp(\sigma^2)+2\right]\sqrt{\exp(\sigma^2)-1}$ (Johnson et al., 1994). This equation can be used to assess the degree of departure from normality on the original scale under the assumed normal model on the logarithmic scale.

THE DATASET

We consider a sunflower experiment for evaluation of 25 hybrid cultivars and 5 checks carried out at three locations across southern and central Spain in the year 2013. The trial was designed as 5 × 6 row-column design of 30 entries with 5 checks. The number of replications was six for two locations (AR01, MO01) and three for location CB01. Locations with six and three replicates had 2-row and 4-row plots, respectively. Dry yield (DY) and oil content (OC) were directly measured from harvested seed of each plot. DY was corrected for moisture and OC was obtained employing NMR (nuclear magnetic resonance) spectroscopy on a subsample from each plot. The field trial AR01 was irrigated. The number of plots with data on both component traits or missing data on one or both traits (before residual analysis) is shown in Table 2. We inspected studentized residuals based on univariate analysis of the component traits and discarded any observation with an absolute residual greater than two. This threshold is rather stringent and it was chosen to generate further imbalance in the data. The corresponding numbers of observations are shown in brackets in Table 2.

Table 2. Number of plots in each trial with data on both component traits, on only one trait, or on none of the component traits. In brackets: Number of plots after deleting observations with absolute studentized residual larger than two.

	Trial (location)		
Traits present	AR01	CB01	MO01
OC & DY	175 (151)	90 (87)	170 (158)
OC only	5 (16)	0 (1)	4 (12)
DY only	0 (10)	0 (2)	0 (4)
None	0 (3)	0 (0)	6 (6)

OC = oil content; DY = dry yield.

The factors used for analysis are shown in Table 3.

Table 3. Factors used for coding the mixed models for analysis of the sunflower data.

Factor	Description
Gen	Genotype
Trl	Trial (location)
Rep	Replicate within trial
Row	Row within replicate
Col	Column within replicate
Trait	Component trait; levels OC and DY

The model for a single trait is:

$$\begin{aligned} Y &= Gen + Gen.Trl + Trl/Rep/(Row \times Col) \\ &= Gen + Trl : Gen.Trl + Trl.Rep + Trl.Rep.Row + Trl.Rep.Col + \\ &+ \underline{Trl.Rep.Row.Col} \end{aligned} \quad (12)$$

We take the trial main effect (Trl) as fixed because there are only three trials and inter-trial information is usually small (Piepho and Möhring 2006). For multiple traits, this is expanded as follows (Piepho et al. 2004, Piepho and Möhring 2011):

$$\begin{aligned} Y &= Gen + Trait + Gen.Trait + Trl + Trl.Trait : \boldsymbol{Gen.Trl.}\mathbf{Trait} \\ &+ \boldsymbol{Trl.Rep.}\mathbf{Trait} + \boldsymbol{Trl.Rep.Row.}\mathbf{Trait} + \boldsymbol{Trl.Rep.Col.}\mathbf{Trait} + \\ &\quad \underline{\boldsymbol{Trl.Rep.Row.Col.}\mathbf{Trait}} \end{aligned} \quad (13)$$

The SAS code for fitting the multi-trait model (13) is given in Box 3. The corresponding code for GenStat is given in Box 4 in the Appendix.

Box 3. SAS code for fitting the multivariate model (13) to sunflower data and computing genotype means across traits.

```
proc mixed;
class trait trl rep row col gen;
model log_Y=trait|gen trl trl*trait;
random trait / subject=trl*rep type=unr;
random trait / subject=trl*gen type=unr;
random trait / subject=trl*rep*row type=unr;
random trait / subject=trl*rep*col type=unr;
repeated trait / subject=trl*rep*row*col type=unr;
lsmeans gen / cov;
run;
```

TROUBLESHOOTING WHEN FITTING THE MULTIVARIATE MODEL

Convergence problems are not uncommon with multivariate mixed models. We observed different behaviour of the three SAS procedures MIXED, GLIMMIX and HPMIXED and strong dependence of convergence behaviour on good starting values. The following three-stage strategy was found to work reasonably well using the MIXED procedure and the UNR structure:

(1) Fit univariate models to the individual component traits.

(2) Fit a multivariate model, fixing the trait-specific variance components at values obtained in stage (1). Thus, only the correlations are estimated at this stage.

(3) Re-estimate all variance-covariance parameters (variances and correlations) of the multivariate model, using the estimates of variances from stage (1) and the correlation estimates from stage (2) as starting values.

Another option is as follows:

(1) Fit univariate models to the individual component traits $\left(\tilde{X}^{(1)}, \tilde{X}^{(2)}, ...\right)$.
(2) Fit a univariate model to pairwise sums of individual component traits.
(3) Compute covariances (and correlations) between corresponding effects of component traits based on the fact that the equation

$$\text{var}\left[\tilde{X}^{(1)} + \tilde{X}^{(2)}\right] = \text{var}\left(\tilde{X}^{(1)}\right) + \text{var}\left(\tilde{X}^{(2)}\right) + 2\,\text{cov}\left(\tilde{X}^{(1)}, \tilde{X}^{(2)}\right) \Leftrightarrow$$
$$\text{cov}\left(\tilde{X}^{(1)}, \tilde{X}^{(2)}\right) = \frac{\text{var}\left[\tilde{Y}\right] - \text{var}\left(\tilde{X}^{(1)}\right) - \text{var}\left(\tilde{X}^{(2)}\right)}{2}$$

also holds for the individual random effects.
(4) Use variances from (1) and covariances from (3) as starting values for multivariate analysis and re-estimate all variance-covariance parameters (variances and correlations).
The first option was used in the example.

RESULTS

Residual plots before removal of outliers are shown in Figures 1, 2 and 3. Overall the logarithmic transformation as well as the original data yield reasonable residual plots. In the plots for DY and DOY there are a few observations that appear to be slightly outlying, both on the original scale and slightly more so on the logarithmic scale, but the overall impression of all of these plots is largely inconspicuous. Thus, an analysis on a logarithmic scale seems reasonable for most practical purposes. As explained before, for further analysis of data on the logarithmic scale we removed any observation that had an absolute studentized residual larger than two based on a univariate analysis for the component trait under consideration based on a fit of model (12) with fixed genotype main effect.

RESIDUAL PLOTS FOR UNTRANSFORMED AND LOG-TRANSFORMED DATA

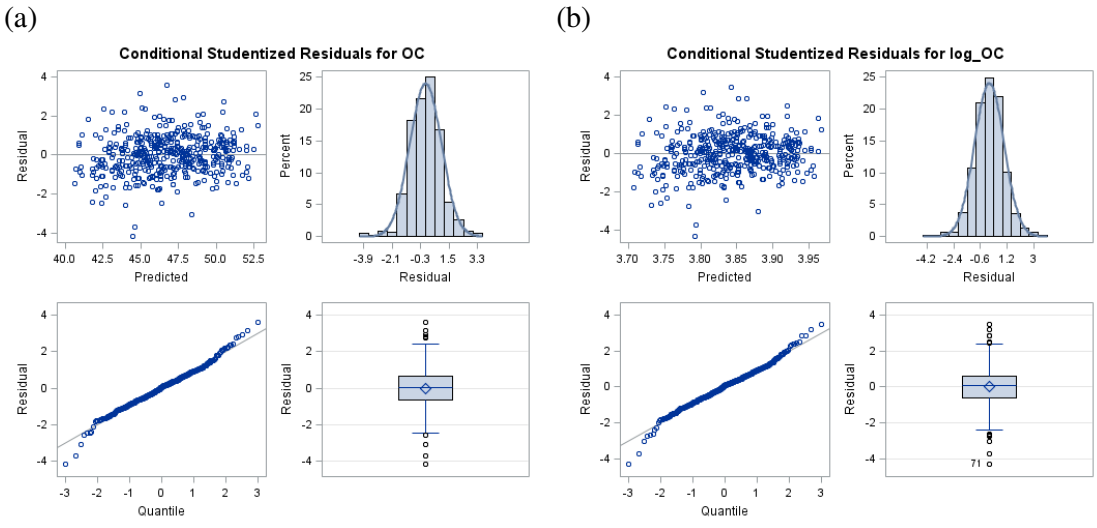

Figure 1. Conditional studentized residual plots for analysis of (a) oil content (OC) and (b) natural logarithm of OC. The plots are based on complete data.

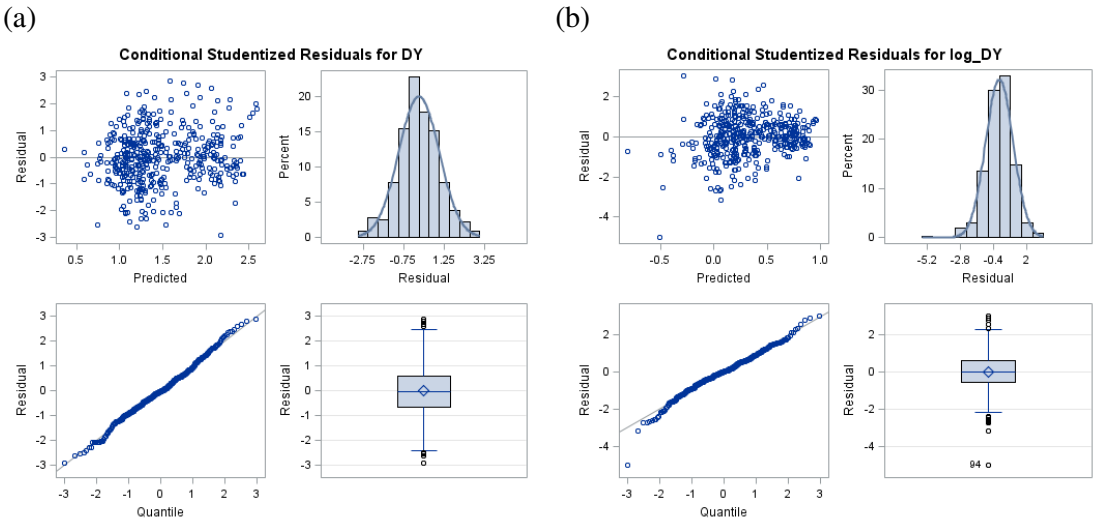

Figure 2. Conditional studentized residual plots for analysis of (a) kernel dry yield (DY) and (b) natural logarithm of DY. The plots are based on complete data.

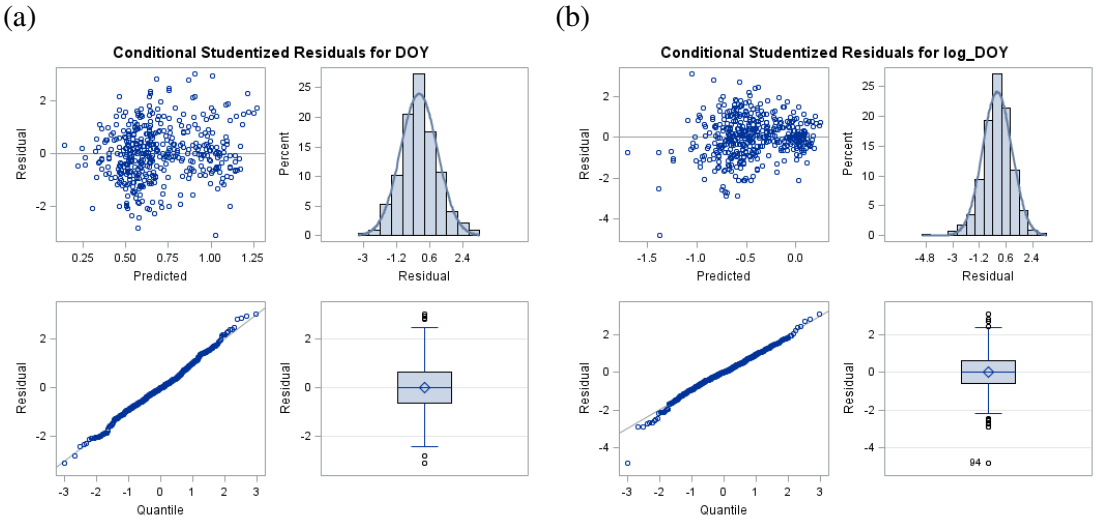

Figure 3. Conditional studentized residual plots for analysis of (a) dry oil yield (DOY) and (b) natural logarithm of DOY. The plots are based on complete data.

CORRELATIONS AMONG ADJUSTED MEANS

We computed means for the complex trait using different models (bivariate and univariate) and data (all data and complete plots only) on both the logarithmic and the original scales. Overall, the correlations are quite high (Table 4). The lowest correlations occurred when means were computed from different data (all versus complete plots only).

Table 4. Correlation (Pearson above diagonal, Spearman below diagonal) among genotype means for DOY obtained by different methods and models.

	Biv_log_cp	Biv_back_cp	Biv_log_all	Biv_back_all	Uni_original
Biv_log_cp	1	0.998	0.993	0.991	0.976
Biv_back_cp	1	1	0.990	0.992	0.977
Biv_log_all	0.984	0.984	1	0.998	0.970
Biv_back_all	0.984	0.984	1	1	0.971
Uni_original	0.965	0.965	0.949	0.949	1

biv = bivariate analysis, log = log-transformed data, cp=complete-plot data, back = back-transformed means, all = all data, including plots with missing data on one component trait, uni = univariate (original scale).

HERITABILITY

For all models, we computed the heritability using the *ad hoc* method of Piepho and Möhring (2007) as $H^2 = \sigma_g^2 / \left(\sigma_g^2 + 0.5\bar{v}_d\right)$, where $\bar{v}_d$ is the average variance of a difference among adjusted genotype means. To estimate $\bar{v}_d$, the models were fitted taking the genotype main effect as fixed (Tables 5 and 6), whereas estimating the genotypic variance required taking this effect as random (Tables 7 and 8). We also estimated skewness on the original scale for models assuming normality on the log-scale. Univariate analyses of the component traits were used to obtain starting values for the bivariate analyses (Table 9).

Table 5. Estimates of variance parameters for bivariate models (component traits, log-transformed data) fitted to (i) complete plot data and all data. Genotype and trial main effects fixed. Var = variance, Corr = correlation.

Covariance parameter	Subject effect	Variance parameter estimate	
		Complete plots only	All data
Var(DY)	Trial.Gen	0.0106	0.0108
Var(OC)	Trial.Gen	0.000195	0.000238
Corr(DY,OC)	Trial.Gen	0.377	0.389
Var(DY)	Trial.Rep	0.0255	0.0269
Var(OC)	Trial.Rep	0.000461	0.000441
Corr(DY,OC)	Trial.Rep	0.169	0.181
Var(DY)	Trial.Rep.Row	0.00127	0.000939
Var(OC)	Trial.Rep.Row	0.000108	0.000086
Corr(DY,OC)	Trial.Rep.Row	0.348	0.354
Var(DY)	Trial.Rep.Col	0.0138	0.0133
Var(OC)	Trial.Rep.Col	0.000144	0.000183
Corr(DY,OC)	Trial.Rep.Col	0.417	0.412
Var(DY)	Trial.Rep.Row.Col	0.0140	0.0146
Var(OC)	Trial.Rep.Row.Col	0.000409	0.000425
Corr(DY,OC)	Trial.Rep.Row.Col	0.134	0.136

Table 6. Variance components for random effects of DOY. Direct: Analysis of DOY or log(DOY) values per plot. Indirect: Derived from bivariate model. Genotype and trial main effects fixed.

Effect/parameter	DOY (direct)	log(DOY) (direct)	log(DOY) (indirect, complete plots only)	log(DOY) (indirect, all plots)
Trial.Gen	0.00533	0.0121	0.0118	0.0122
Trial.Rep	0.00988	0.0270	0.0272	0.0286
Trial.Rep.Row	0.000723	0.00163	0.00164	0.00123
Trial.Rep.Col	0.00400	0.0151	0.0151	0.0147
Trial.Rep.Row.Col	0.00680	0.0150	0.0151	0.0157
Total variance		0.0709	0.0708	0.0724
Skewness		0.833	0.832	0.843

Table 7. Estimates of variance parameters for bivariate models (component traits, log-transformed data) fitted to (i) complete plot data and (ii) all data. Trial main effects fixed. Var = variance, Corr = correlation.

Covariance parameter	Subject effect	Variance parameter estimate	
		Complete plots only	All data
Var(DY)	Gen	0.00307	0.00318
Var(OC)	Gen	0.00226	0.00216
Corr(DY,OC)	Gen	-0.0277	-0.114
Var(DY)	Trial.Gen	0.0111	0.0112
Var(OC)	Trial.Gen	0.000199	0.000241
Corr(DY,OC)	Trial.Gen	0.382	0.402
Var(DY)	Trial.Rep	0.0254	0.0269
Var(OC)	Trial.Rep	0.000461	0.000441
Corr(DY,OC)	Trial.Rep	0.170	0.180
Var(DY)	Trial.Rep.Row	0.00127	0.000883
Var(OC)	Trial.Rep.Row	0.000104	0.000083
Corr(DY,OC)	Trial.Rep.Row	0.327	0.311
Var(DY)	Trial.Rep.Col	0.0135	0.0130
Var(OC)	Trial.Rep.Col	0.000144	0.000182
Corr(DY,OC)	Trial.Rep.Col	0.407	0.405
Var(DY)	Trial.Rep.Row.Col	0.0140	0.0146
Var(OC)	Trial.Rep.Row.Col	0.000411	0.000426
Corr(DY,OC)	Trial.Rep.Row.Col	0.138	0.143

Implied skewness on the original scale was moderate for all analyses (Tables 6 and 8). The numerical differences in heritability should not be over-interpreted because of the very small number of genotypes. To gain some insight with this dataset, we fixed variance-covariance parameters at values obtained from an analysis of all data, taking the genotype main effect as random (Table 7). These fixed values were then used to estimate heritability based on bivariate analyses of all data and based on complete plots only (Table 10). As expected, heritability is slightly larger when all data are used, as opposed to using data only from plots with complete data on both component traits (Table 1). The difference in heritability is not dramatic, but the result does show that some improvement is possible by analysing all available data.

Table 8. Variance components for random effects of DOY. Direct: Analysis of DOY or log(DOY) values per plot. Indirect: Derived from bivariate model. Trial main effects fixed.

Effect/parameter	DOY (direct)	log(DOY) (direct)	Log(DOY) (indirect, complete plots only)	Log(DOY) (indirect, all data)
Gen	0.00215	0.00516	0.00518	0.00474
Trial.Gen	0.00530	0.0125	0.0124	0.0128
Trial.Rep	0.00989	0.0270	0.0270	0.0286
Trial.Rep.Row	0.000705	0.00161	0.00162	0.00114
Trial.Rep.Col	0.00389	0.0150	0.0147	0.0144
Trial.Rep.Row.Col	0.00685	0.0150	0.0151	0.0158
Total variance	0.0288	0.0763	0.0760	0.0774
Skewness	-	0.867	0.865	0.874
Heritability (*ad hoc*)	0.459	0.473	0.479	0.454

Table 9. Variance components for random effects of log(OC) and log(DY) based on univariate analyses. Trial main effects fixed. Genotype main effect fixed or random.

Effect/parameter	log(OC)		log(DY)	
	Gen fixed	Gen random	Gen fixed	Gen random
Gen		0.00218		0.00328
Trial.Gen	0.000236	0.000239	0.0105	0.0109
Trial.Rep	0.000449	0.000448	0.0266	0.0266
Trial.Rep.Row	0.000087	0.000085	0.00100	0.000974
Trial.Rep.Col	0.000173	0.000173	0.0133	0.0131
Trial.Rep.Row.Col	0.000429	0.000429	0.0145	0.0146

Table 10. Heritability (H^2) and mean variance of a difference ($\bar{v}_d$) for DOY based on bivariate analysis of log(DY) and log(OC) using variance parameter estimates obtained from fit of all data. Comparison of the results when complete plots only are used versus use of all data.

Effect/parameter	Complete plots only	All data
Genetic variance	0.00474	0.00474
Mean variance of a difference	0.0120	0.0118
Heritability (*ad hoc*)	0.441	0.446

DISCUSSION

This paper has shown, using a real example, that fitting a multivariate linear mixed model to the component traits (e.g. dry matter yield and oil content) on the logarithmic scale allows making full use of all the information on the complex trait (e.g. dry oil yield) when some of the data on the component traits are missing.

The complex trait is an exact product of the component traits. For clarity, it is worth pointing out that under our multiplicative model there is no independent error associated with the complex trait in the sense that all error terms in this model stem from the errors of the component traits.

In this paper we have focussed on the case of a complex trait that is computed from component traits. However, there are also applications, where a component trait is computed

from a complex trait and one or several component traits. For example, kernel weight in cereals $\left(X^{(1)}\right)$ is often calculated based on grain yield per unit area $\left(Y\right)$, number of spikes per unit area $\left(X^{(2)}\right)$, and number of kernels per spike $\left(X^{(3)}\right)$ as $X^{(1)} = Y/\left(X^{(2)}X^{(3)}\right)$. With minor modifications, the multivariate approach proposed in this paper is also applicable in this scenario. For the example of kernel weight, we would fit a multivariate mixed model to $\left(\tilde{Y} = \log\left(Y\right), \tilde{X}^{(2)} = \log\left(X^{(2)}\right), \tilde{X}^{(3)} = \log\left(X^{(3)}\right)\right)$ and then make inferences on kernel weight on the log scale based on $\tilde{X}^{(1)} = \log\left(X^{(1)}\right) = \tilde{Y} - \tilde{X}^{(2)} - \tilde{X}^{(3)}$.

In implementing our mixed models, we have favoured REML over competing methods. REML has become the standard method for fitting mixed models because this method has several desirable properties. For example, variance component estimates tend to be less biased than full ML estimates, and REML estimates are consistent (as are ML estimates) (Searle et al. 1992).

One might be inclined to analyse each component trait separately, computing genotype means for each trait, and then estimating the mean of the complex trait simply by multiplying the corresponding genotype means of the component traits. But this analysis is bound to produce biased results when the component traits are correlated. It is also in disagreement with our multiplicative model. To see this, consider the two component trait values $X^{(1)}$ and $X^{(2)}$. We are interested in estimating the "mean" of the product of the component trait values, $Y = X^{(1)} \times X^{(2)}$, i.e., the expected value $E\left(Y\right) = E\left(X^{(1)}X^{(2)}\right)$. From the definition of a covariance (Rice 1995), we have

$$E\left(X^{(1)}X^{(2)}\right) = E\left(X^{(1)}\right)E\left(X^{(2)}\right) + COV\left(X^{(1)}, X^{(2)}\right), \tag{14}$$

where $COV\left(X^{(1)}, X^{(2)}\right)$ denotes the covariance of $X^{(1)}$ and $X^{(2)}$. So obviously, the simple plug-in approach produces a bias when $COV\left(X^{(1)}, X^{(2)}\right) \neq 0$, which will be the rule rather than the exception in yield component analysis (Fraser and Eeaton 1983, Piepho 1995, Spaarnaij and Bos 1993, also see Tables 5 and 7).

Analysing yield data on a logarithmic scale is not uncommon. One prominent example is the seminal paper by Finlay and Wilkinson (1963), who performed their proposed regression for stability assessment in order to stabilize the variance and better meet the assumption of linearity. Also, many breeders compute relative yields compared to check varieties in order to analyse their trials (Piepho 1994, Yau and Hamblin 1994, Schwarzbach et al. 2007). Use of relative yields also implies a multiplicative model for original yields, rather than an additive model. Limpert et al. (2001) provide a very lucid review of diverse examples from biology in which data more closely follow a log-normal rather than a normal distribution. Francis Galton (cited in Lynch and Walsh 1998, p.295) was the first to point out that taking logarithms is expected to achieve approximate normality for traits that can be written as products of component traits (see eq. 3), due to the resulting transformation to an additive relation (eq. 4) and operation of the central limit theorem, and he put this fact forward to explain the commonness of the log-normal distribution. It should also be noted, however, that the log-normal distribution approaches a normal distribution when the variance on the logarithmic scale becomes small, so both distributional assumptions may hold approximately for a given dataset. For yield component analysis, working on the logarithmic scale as shown here is mainly a matter of convenience because the analysis can proceed by linear inference as opposed to the more complex analysis based on the multiplicative model that holds on the original scale (Brown and Alexander 1991).

We also mention here that analysis of yield components on a logarithmic scale bears some resemblance to compositional data analysis (Aitchison 1986, Pawlowsky-Glahn and Buccianti 2011). Compositional data arise when the total quantity of some trait is decomposed into its components. For example, soil samples may be decomposed into

components defined by particle size, i.e., clay, silt and sand fractions. If these components are expressed as fractions of the total, the data are referred to as compositional. Yield components on the logarithmic scale could be regarded as compositional data if expressed as a fraction of the total, corresponding to logarithmic yield. As yet, there does not seem to be a published application of methods for compositional data to yield components. For a detailed discussion of the relation between yield component analysis and compositional data (i.e., composite variables or composite scores) see Kozak (2010).

For the analysis by a multiplicative model as proposed here it is crucial that all component traits are assessed at the plot level. It may also be mentioned that dry matter yield itself is computed as the product of fresh matter yield and dry matter content. Sometimes, for the sake of simplicity, dry matter content is computed from a pooled sample that comprises samples from different plots. Subsequently, dry matter yield for each plot is computed as the product of fresh matter yield from the plot and the dry matter content value obtained from the pooled sample. It is important to note, however, that the dry matter measurement for the pooled sample does not capture the between-plot variance for dry matter content and hence a valid statistical analysis based on plot dry matter yield data computed this way is not possible.

REFERENCES

Aitchison J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, London.

Brown, D., Alexander, N. (1991). The analysis of variance and covariance of products. *Biometrics* 47, 429–444.

Finlay K.W., Wilkinson G.N. (1963). The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research* 14, 742–754.

Fraser J., Eaton G.W. (1983). Application of yield component analysis to crop research. *Field Crops Abstracts* 36, 787–797.

Gołaszewski J. (1996). A method of yield component analysis. *Biometrical Letters* 33, 79–88.

Johnson N.L., Kotz S., Balakrishnan N. (1994). *Continuous univariate distributions*. Wiley, New York.

Kozak M. (2004). New concept of yield component analysis. Biometrical Letters 41, 59-69.

Kozak M. (2010). Note on two methods of additive yield component analysis. *Biometrical Letters* 47, 129–132.

Kozak M., Mądry W. (2006). Note on yield component analysis. *Cereal Research Communications* 34, 933–940.

Limpert E., Stahel W.A., Abbt, M. (2001). Log-normal distributions across the sciences. Keys and clues. *Bioscience* 51, 341–352.

Lynch M., Walsh B. (1998). Genetics and analysis of quantitative traits. Sinauer, Sunderland.

Pawlowsky-Glahn V., Buccianti A. (2011). *Compositional data analysis. Theory and applications.* Wiley, New York.

Piepho H.P. (1994). A comparison of the ecovalence and the variance of relative yield as measures of stability. *Journal of Agronomy and Crop Science* 173, 1–4. (Erratum: (1995) 174, 216).

Piepho H.P. (1995). A simple procedure for yield component analysis. *Euphytica* 84, 43–48.

Piepho H.P. (2009). Data transformation in statistical analysis of field trials with changing treatment variance. *Agronomy Journal* 101, 865–869.

Piepho H.P., Büchse A., Emrich K. (2003). A hitchhiker's guide to the mixed model analysis of randomized experiments. *Journal of Agronomy and Crop Science* 189, 310–322.

Piepho H.P., Möhring J. (2006). Selection in cultivar trials – is it ignorable? *Crop Science* 46, 192–201.

Piepho H.P., Möhring J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177, 1881–1888.

Piepho H.P., Möhring J. (2011). On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS System. *Crop Science* 51, 2449–2454.

Rice J.A. (1995). *Mathematical statistics and data analysis. Second edition*. Duxbury Press, Belmont.

Schwarzbach E., Hartmann J., Piepho H.P. (2007). Multiplicative main cultivar effects in Czech official winter wheat trials 1976-2005. *Czech Journal of Genetics and Plant Breeding* 43, 117–124.

Searle S.R., Casella G., McCulloch C.E. (1992). *Variance components*. Wiley, New York.

Sparnaaij L.D., Bos I. (1993). Component analysis of complex characters in plant breeding. I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character. *Euphytica* 70, 225–235.

Verbeke G., Molenberghs G. (1997). *Linear mixed models for longitudinal data.* Springer, Berlin.

Yau S.K., Hamblin J. (1994). Relative yield as a measure of entry performance in variable environments. *Crop Science* 34, 813–817.

APPENDIX

The GenStat code for fitting the multi-trait model (13) is given in Box 4. In difference to SAS, GenStat requires the data for different traits from the same plot to be in a single record. A further difference is that as a default, GenStat uses a product-type parameterization of the variance-covariance structure, which factors out a single residual variance component. Hence, all parameter estimates are, in fact, ratios of a variance component relative to that residual. Multiplying these ratios (called "gammas") by the residual (called "Sigma2") produces the variance and covariance estimates corresponding to those obtained with SAS. There is also an option in the REML directive to switch to the covariance parameterization (referred to as "sigmas").

Box 4. Genstat code for fitting the multivariate model (13) to sunflower data and computing genotype means across traits.

```
VCOMP [FIXED=%_variable/(%_Gen + %_Trl); FACTORIAL=9;
CONST=omit] (%_Trl.%_Gen + %_Trl.%_Rep + %_Trl.%_Rep.%_Row +
%_Trl.%_Rep.%_Col).%_variable + %_units.%_variable;

VSTRUCTURE [%_Trl.%_Gen.%_variable] FACTOR=%_variable; MODEL=unstructured;
VSTRUCTURE [%_Trl.%_Rep.%_variable] FACTOR=%_variable; MODEL=unstructured;
VSTRUCTURE [%_Trl.%_Rep.%_Row.%_variable; MODEL= unstructured;
VSTRUCTURE [%_Trl.%_Rep.%_Col.%_variable] FACTOR=%_variable; MODEL=
unstructured;
VSTRUCTURE [%_units.%_variable] FACTOR=%_variable; MODEL=unstructured;
REML [PRINT=model,components,waldTests, means; MAXCYCLE=20;
FMETHOD=automatic; PSE=differences; MVINCLUDE=explanatory,yvariate;
METHOD=Fisher] _Data;
```