Annex VIII

Analysis of data of Annex VII with SAS/STAT® 15.1.

Germination indices are analysed (see Annex VII for the dataset used) as an example.

First, homogeneity of variances is tested with the GLM procedure. Heteroskedasticity can be expected for these indices when they correspond to largely different time-courses, which anyway is not the case here.

```
proc GLM;
class temp;
model MGT MGR CUG = temp;
means temp / Tukey hovtest=Levene;
run;
```

RESULTS (excerpts):

Table 1. MGT.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.23647245	0.11823622	38.99	<.0001
Error	15	0.04548365	0.00303224		
Corrected Total	17	0.28195610			

Table 2. MGT.

R-Square	Coeff Var	Root MSE	MGT Mean
0.838685	1.781901	0.055066	3.090284

Figure 1.

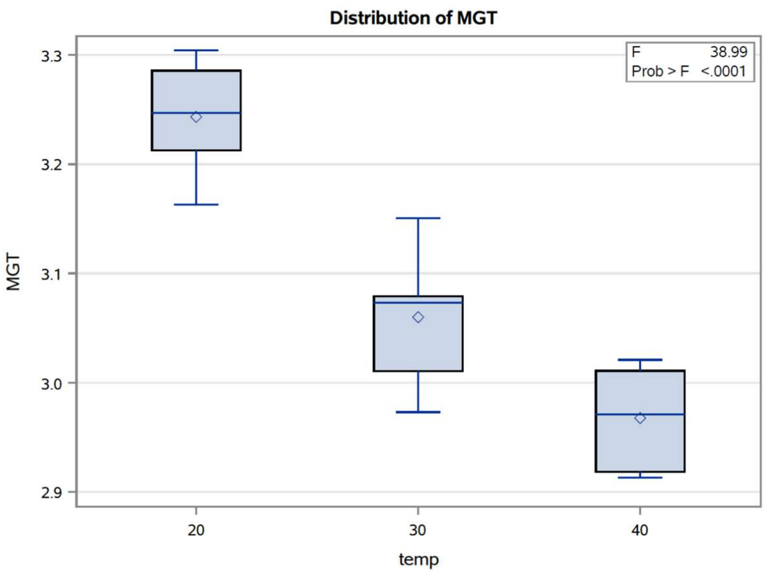


Table 3. MGR.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.00253526	0.00126763	37.56	<.0001
Error	15	0.00050631	0.00003375		
Corrected Total	17	0.00304157			

Table 4. MGR.

R-Square	Coeff Var	Root MSE	MGR Mean
0.833537	1.792481	0.005810	0.324121

Figure 2.

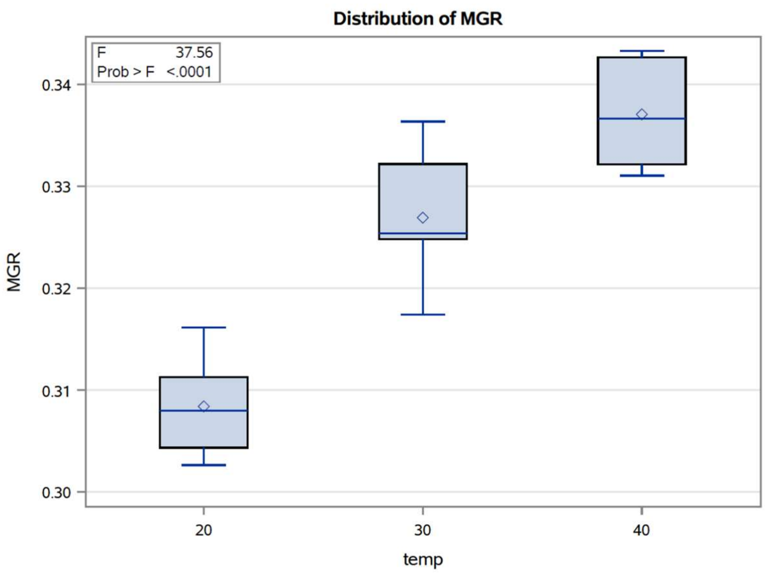


Table 5. CUG.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9.22626284	4.61313142	5.64	0.0149
Error	15	12.26068153	0.81737877		
Corrected Total	17	21.48694437			

Table 6. CUG.

R-Square	Coeff Var	Root MSE	CUG Mean
0.429389	27.79870	0.904090	3.252275

Figure 3.

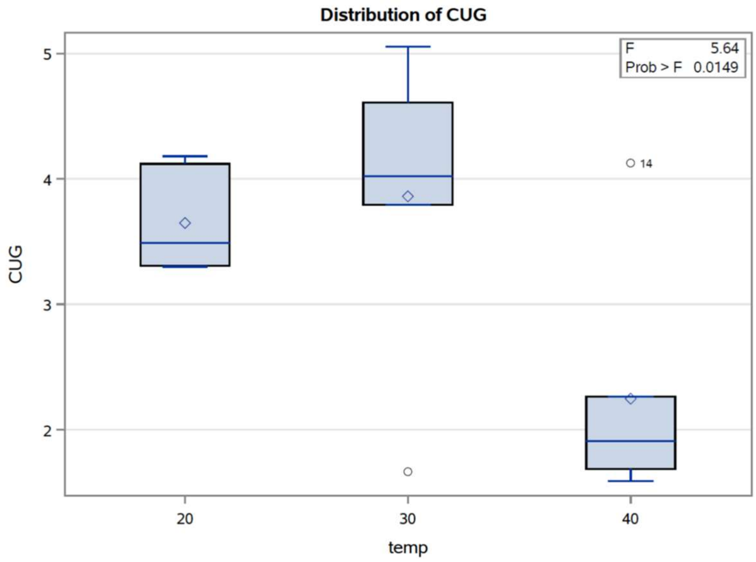


Table 7. MGT.

Levene's Test for Homogeneity of MGT Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
temp	2	3.857E-6	1.929E-6	0.29	0.7546
Error	15	0.000101	6.723E-6		

Table 8. MGR.

Levene's Test for Homogeneity of MGR Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
temp	2	6.72E-10	3.36E-10	0.42	0.6638
Error	15	1.196E-8	7.97E-10		

Table 9. CUG.

Levene's Test for Homogeneity of CUG Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
temp	2	3.1483	1.5742	0.87	0.4381
Error	15	27.0689	1.8046		

Figure 4.

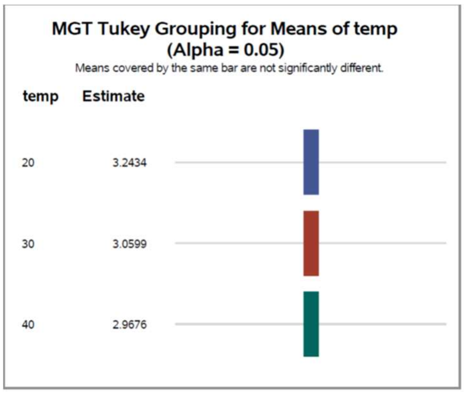


Figure 5.

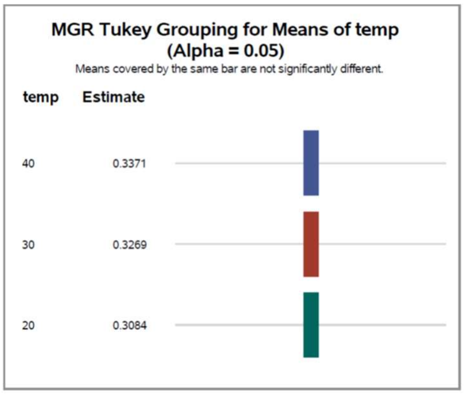
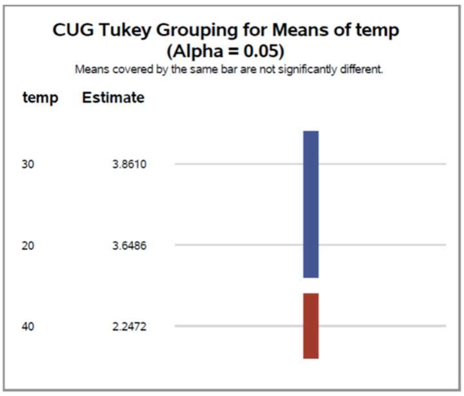


Figure 6.



The modelled factor, 'temp', has a significant ($P < 0.0001$) effect on MGT (Table 1), with a high determination coefficient (R-square) and a tiny coefficient of variation (Table 2). The distribution of the data does not show anomalies (Figure 1). MGR, being the reciprocal of MGT, has the same statistical features (Tables 3 and 4), and the distribution of the data is a mirror image of MGT (Figure 2). The modelled 'temp' factor affects CUG too, though in this case it is somewhat less significant ($P = 0.0149$; Table 5), with a low determination coefficient (R-square) and a very high coefficient of variation (Table 6). Figure 3 illustrates the strong variability of CUG among replicated plates, essentially due to single outliers. Clearly, this index is much more sensitive to stochastic variations among replicates than MGT and MGR. Nonetheless, both MGT, MGR, as well as CUG do not appear to display heterogeneous variances in this instance (Tables 7-9). Tukey's tests for mean comparisons indicate a significant difference among all three temperature levels for both MGT (Figure

4) and MGR (Figure 5), but only between 40 °C and the two other temperatures in the case of CUG (Figure 6).

The statistical analyses of MGT and MGR appear solid, and, given the small plate effect (see Annex IV for general considerations on this topic) and the absence of heteroskedasticity, no further analysis is considered for these data. This might well hold for CUG too, but given the much lower fit of its model, and for the sake of illustrating the use of the GLIMMIX procedure in case of heteroskedastic data, a conditional model is additionally considered:

```
/*conditional model*/
proc GLIMMIX method=Laplace;
class temp plate;
model CUG = temp;
random intercept / subject=plate(temp) group=temp;
lsmeans temp / cl plot=meanplot adjust=Tukey;
covtest zeroG;
covtest homogeneity;
run;
```

RESULTS (excerpts):

Table 10.

Model Information	
Data Set	WORK.REFFILE
Response Variable	CUG
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix Blocked By	plate(temp)
Estimation Technique	Maximum Likelihood
Likelihood Approximation	Laplace
Degrees of Freedom Method	Containment

Table 11.

Dimensions	
G-side Cov. Parameters	3
R-side Cov. Parameters	1
Columns in X	4
Columns in Z per Subject	3
Subjects (Blocks in V)	18
Max Obs per Subject	1

Table 12.

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	7
Lower Boundaries	4
Upper Boundaries	0
Fixed Effects	Not Profiled
Starting From	GLM estimates

Table 13.

Fit Statistics for Conditional Distribution	
-2 log L(CUG | r. effects)	0.98
Pearson Chi-Square	0.77
Pearson Chi-Square / DF	0.04

Table 14.

Covariance Parameter Estimates				
Cov Parm	**Subject**	**Group**	**Estimate**	**Standard Error**
Intercept	plate(temp)	temp 20	0.02096	0.07928
Intercept	plate(temp)	temp 30	1.0381	0.6665
Intercept	plate(temp)	temp 40	0.6354	0.4340
Residual			0.1163	.

Table 15.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
temp	2	15	7.06	0.0069

Table 16.

temp Least Squares Means								
temp	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
20	3.6486	0.1513	15	24.12	<.0001	0.05	3.3262	3.9711
30	3.8610	0.4386	15	8.80	<.0001	0.05	2.9260	4.7959
40	2.2472	0.3540	15	6.35	<.0001	0.05	1.4928	3.0017

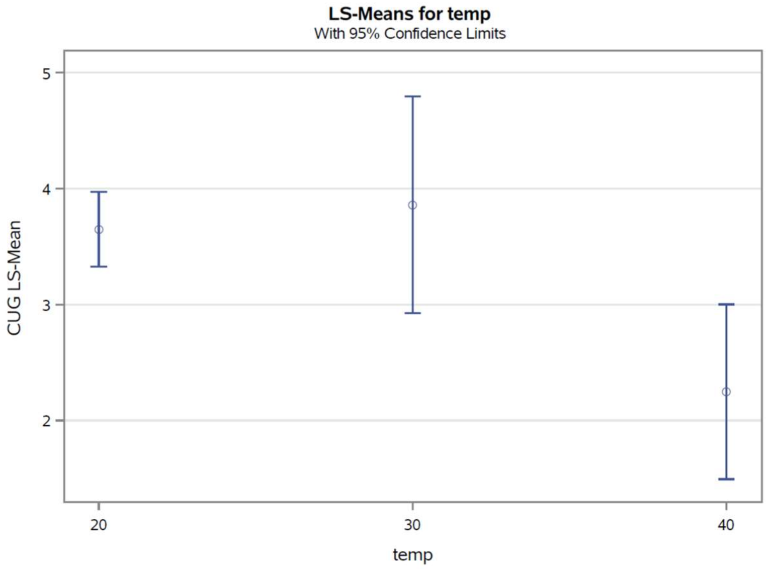


Figure 7.

Table 17.

Differences of temp Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
temp	temp	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
20	30	-0.2123	0.4640	15	-0.46	0.6538	0.8918	0.05	-1.2013	0.7767	-1.4175	0.9929
20	40	1.4014	0.3849	15	3.64	0.0024	0.0064	0.05	0.5809	2.2219	0.4016	2.4013
30	40	1.6137	0.5636	15	2.86	0.0118	0.0300	0.05	0.4123	2.8151	0.1497	3.0778

Table 18.

Tests of Covariance Parameters Based on the Likelihood					
Label	DF	-2 Log Like	ChiSq	Pr > ChiSq	Note
No G-side effects	3	44.1703	5.85	0.0408	MI
Homogeneity	2	44.1703	5.85	0.0536	DF

DF: P-value based on a chi-square with DF degrees of freedom.
MI: P-value based on a mixture of chi-squares.

As the event/trial syntax was not used, the response variable is assumed to have a Gaussian distribution and Maximum Likelihood (ML) is used as estimation technique (Table 10) since the Laplace approximation was invoked. Table 11 shows that three G-side variance/covariance parameters were computed, namely, the three between-plates variances separately estimated for each temperature level, as required by the ‘group=’ option. One R-side parameter was estimated as well, i.e. the residual scale parameter, though no ‘random residual’ statement was present, because for mixed models with normal data, the GLIMMIX procedure computes, from the initial GLM estimates, the scale parameter, which, for Gaussian data, represents the maximum likelihood estimate of the error variance, and not an overdispersion parameter. The X matrix (the matrix for the fixed effects) comprises 4 columns, corresponding to a column for the intercept and three columns for the levels of the ‘temp’ effect. There are 3 columns in the Z matrix (the matrix for G-side random effects) for this model, corresponding to the three distinct variances modelled for the three temperature levels. On the other hand, there are 18 subjects (Blocks in V, where the V matrix includes all the random effects, both on the G-side and R-side), since data are processed by subjects, that is, the V matrix is subdivided in as many blocks as the number of subjects; in fact, Table 10 already indicated the Variance Matrix blocked by 'plate(temp)'. Note that data are processed by subjects because these are explicitly called in the 'random' statement by the 'subject=plate(temp)' option as subjects modelled in terms of random intercepts; if the same statement had been formulated with a shorter syntax (to wit, 'random plate(temp) / ...') the whole matrix would have been considered a single block (but the estimates would have been unaffected). Seven parameters are under optimization (Table 12): the means of the three temperature levels (the fixed effects), their three variances (the random effects), and the scale parameter. The means have no estimation boundaries, whereas the random effects parameters have a lower boundary of zero. The Pearson Chi-Square / DF parameter (which, for a conditional model with Gaussian data, does not represent an estimation of the scale parameter, but, rather, another measure of variability of the observations around the mean) indicates that there is little residual variation (Table 13), and, indeed, the residual variance is low (Table 14). The estimates of the random variances among plates at the diverse temperatures (Table 14) suggest that the variance among plates at 20 °C was smaller, though this difference is not significant (Table 9), because larger variances at 30 and 40 °C were essentially due to a single outlier datum in each case (Figure 3). This model indicates a stronger significance of the temperature effect on CUG (P = 0.0069; Table 15) than found by the GLM analysis (P = 0.0149; Table 5). This is presumably owing to the inclusion of the random effects of plates into the model, which leaves out only a small residual variance (Tables 13 and 14), and therefore improves the power of the analysis. In this respect, it ought to be noticed that, without the Laplace approximation (which invokes both ML, instead of REML, and the 'noprofile' option) or just the use of the 'noprofile' option (which requests that the scale parameter be included into the optimization rather than profiled from it) in the GLIMMIX statement, this model would incur in troubles because the modelled random effect is equivalent to the residual variance, and, specifically, the estimate of the profiled scale parameter, i.e. the residual variance, is linearly related, and almost equivalent, to the random variances under optimization, thus that the random effects are estimated in conditions close to over-specification. In fact, the variance of the between-plates effect is modelled as a random factor, and the scale parameter models the variance of the interaction between the plate random effect and the 'temp' fixed factor, but plates are nested within temperature levels and nesting corresponds to an entirely unbalanced interaction. As the variance of the fully unbalanced interaction over the three levels of the fixed factor is equivalent (apart from BLUPs shrinkage) to the variance of the nested random factor across the three levels of the fixed factor, the two variance sources are confounded. Under the normality assumption, the random interaction cannot appear in the linear predictor because it is confounded with the residual variance (Gbur et al., 2012). However, Laplace estimates typically exhibit better asymptotic behaviour and, in this specific case, the inclusion of the scale parameter in the optimization solves the conflict between having to estimate the scale parameter as profiled from optimization while random variances, practically equivalent to the scale parameter, are estimated by optimization. Thus, including the scale parameter into the optimization partially overcomes this problem. Nonetheless, the standard error of the residual estimate cannot be calculated (Table 14), suggesting that a conditional model is indeed at the boundaries of estimability for these data. Table 15 displays the LS-means for the three temperature levels, which coincide with those found with the GLM analysis (Figure 3). Figure 7 illustrates the estimated means with their confidence intervals. Results of multiple comparisons (Table 17) confirm a significant difference of CUG at 40 °C from the other two temperatures, as found with GLM (Figure 6). Finally, including the between-plates random effects in the model has a significant impact (at P = 0.05), and considering heterogeneous variances among temperature levels, albeit non-significant at P = 0.05 (like found by GLM analysis; Table 9), represents a non-entirely useless effect (Table 18). The over-specification of the conditional model can be overcome using a marginal model:

```
proc GLIMMIX; /*marginal model*/
class temp plate;
model CUG = temp;
random residual / group=temp;
lsmeans temp / cl plot=meanplot adjust=Tukey;
covtest homogeneity;
run;
```

From a general perspective, it is worth noticing that in both the ‘random’ (G-side) and ‘random residual’ (R-side) statements, the default variance/covariance structure is the ‘variance components’ (vc), which assigns a distinct variance to each of the specified effects. When a random effect is modelled as a G-side effect with the default ‘vc’ variance/covariance structure, however, the scale parameter is modelled together with the random variances (as seen when discussing Table 11), which means that such a structure is equivalent to modelling residuals, i.e. to an R-side effect, with ‘compound symmetry’ variance/covariance structure, typically arising in presence of nested random effects. Modelling, instead, the R-side effect with the default ‘vc’ structure avoids considering the residuals as an R-side parameter. Thus, the above marginal model with the ‘group=’ option only provides separate modelling of the residual variances for the levels of the ‘temp’ factor, and does not incur in the over-specification of the conditional model, which, in absence of the Laplace approximation causes an overlapping of estimations that leads to the warning that “A linear combination of covariance parameters is confounded with the residual variance”. It might be further noticed that in this example it is not even necessary to specify that ‘plate(temp)’ is the effect for which the residuals are modelled, because there is no other residual effect. If it were specified, it could also be processed by subjects making explicit the option that ‘subject=plate(temp)’ in the ‘random residual’ statement. All these syntax variants would lead to the same inference, as would simply including the ‘residual’ keyword amid the options of the ‘random statement’ used in the conditional model (providing that the ‘method=Laplace’ option were removed from the first statement of that model).

RESULTS (excerpts):

Table 19.

Model Information	
Data Set	WORK.REFFILE
Response Variable	CUG
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

Table 20.

Dimensions	
R-side Cov. Parameters	3
Columns in X	4
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	18

Table 21.

Optimization Information	
Optimization Technique	Dual Quasi-Newton
Parameters in Optimization	3
Lower Boundaries	3
Upper Boundaries	0
Fixed Effects	Profiled
Starting From	Data

Table 22.

Fit Statistics	
-2 Res Log Likelihood	40.04
AIC (smaller is better)	46.04
AICC (smaller is better)	48.22
BIC (smaller is better)	48.17
CAIC (smaller is better)	51.17
HQIC (smaller is better)	46.02
Generalized Chi-Square	15.00
Gener. Chi-Square / DF	1.00

Table 23.

Covariance Parameter Estimates			
Cov Parm	**Group**	**Estimate**	**Standard Error**
Residual (VC)	temp 20	0.1648	0.1042
Residual (VC)	temp 30	1.3853	0.8761
Residual (VC)	temp 40	0.9021	0.5705

Table 24.

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
temp	2	15	5.88	0.0130

Table 25.

Differences of temp Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
temp	temp	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
20	30	-0.2123	0.5083	15	-0.42	0.6821	0.9089	0.05	-1.2957	0.8711	-1.5325	1.1079
20	40	1.4014	0.4217	15	3.32	0.0046	0.0121	0.05	0.5026	2.3002	0.3061	2.4967
30	40	1.6137	0.6174	15	2.61	0.0196	0.0485	0.05	0.2977	2.9298	0.009946	3.2175

Table 26.

Tests of Covariance Parameters Based on the Restricted Likelihood					
Label	DF	-2 Res Log Like	ChiSq	Pr > ChiSq	Note
Homogeneity	2	44.9186	4.88	0.0873	DF

DF: P-value based on a chi-square with DF degrees of freedom.

REML is now the estimation technique (Table 19). No G-side effects, and therefore no columns in Z, are modelled (Table 20). No subject effect was specified and thus data are not processed by subjects. There are only three parameters in optimization (Table 21) because a residual likelihood technique is used to compute the objective function and thus the fixed effects are profiled from the optimization (i.e., they are computed analytically as exact values rather than as asymptotic estimation). The Generalized Chi-Square / DF parameter is displayed (Table 22) as a measure of variability of the observations around the approximated model, which assumes the fixed factors as given (since they are profiled). The three error (or random residual) variance estimates are displayed in Table 23, whereas residual variance is no longer estimated as R-side parameter. Significance of the temperature effect on CUG (P = 0.0130; Table 24) is closer to the value found by the GLM analysis (P = 0.0149; Table 5) than to that found by the conditional model with Laplace approximation (P = 0.0069; Table 15). Inference about the differences between the means collimates with the other models (Table 25). The impact of modelling heterogeneous errors for the three means (Table 26) is similar to that found with the conditional model.

The conditional and marginal models do not introduce any substantial change into the inference previously reached with the general linear model, but they, the latter in particular, furnish an example of how to elaborate values of germination indices if these are heteroskedastic. As seen in Annex IV, germination data typically correspond to hierarchical designs, and modelling them with the aid of the R matrix improves robustness of convergence and optimization. Marginal models can therefore be preferred for the analysis of indices and other response variables with Gaussian distribution, because the aggregate observations, i.e the plates, are subjects nested within each treatment rather than true blocks (that is, replicates of the whole set of experimental contrasts), and this can lead to over-specification of the conditional model.

References

Gbur E.E., Stroup W.W., McCarter K.S., Durham S., Young L.J., Christman M., West M. and Kramer M. (2012). *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. American Society of Agronomy: Madison, WI, USA.