Analysis of covariance in agronomy and crop research

Rong-Cai Yang[1] and Patricia Juskiw[2]

[1] Research and Innovation Division, Alberta Agriculture and Rural Development, #307, 7000-113 Street, Edmonton, Alberta, Canada T6H 5T6 and Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta, Canada T6G 2P5 (e-mail: rong-cai.yang@ualberta.ca); and [2] Research and Innovation Division, Alberta Agriculture and Rural Development, Field Crop Development Centre, 5030 - 50 Street, Lacombe, Alberta, Canada T4L 1W8. Received 11 December 2010, accepted 21 February 2011.

Yang, R.-C. and Juskiw, P. 2011. **Analysis of covariance in agronomy and crop research**. Can. J. Plant Sci. **91**: 621–641. Analysis of covariance (ANCOVA) is a statistical technique that combines the methods of the analysis of variance (ANOVA) and regression analysis. However, ANCOVA is an advanced topic that often appears towards the end of many textbooks, and thus, it is either taught cursorily or ignored completely in many statistics classes. Additionally, many elaborated applications of ANCOVA to agronomy and crop research along with uses of the latest statistical software are rarely described in textbooks or classes. The objectives of this paper are to provide an overview on conventional ANCOVA and to introduce more advanced uses of ANCOVA under mixed models. We describe three elaborate applications including (i) the use of ANCOVA for dissecting dosage responses for different treatments, (ii) stability of treatments across multiple environments and (iii) removal of spatial variation that is not effectively controlled by blocking. These analyses illustrate that ANCOVA is either a simpler analysis or provides more information than conventional statistical methods. We provide a technical appendix (Appendix A) on principles and theory underlying mixed-model analysis of ANCOVA along with SAS programs (Appendix B) for more uses and in-depth understanding of this powerful technique in agronomy and crop research.

Key words: Analysis of covariance, dosage response, mixed models, nearest neighbour analysis, orthogonal polynomials, spatial variability, stability analysis, statistical control of errors

Yang, R.-C. et Juskiw, P. 2011. **Analyse de la covariance en recherche agronomique et agricole**. Can. J. Plant Sci. **91**: 621–641. L'analyse de la covariance (ANCOVA) est une méthode statistique qui combine l'analyse de la variance (ANOVA) et l'analyse de régression. Pourtant, l'ANCOVA est un sujet avancé que den nombreux ouvrages de statistique n'abordent souvent qu'à la fin, donc qui n'est enseigné qu'en diagonale, voire ignoré totalement dans maints cours de statistique. D'autre part, rares sont les ouvrages et les cours qui décrivent les multiples applications évoluées de l'ANCOVA en recherche agronomique et agricole ou les plus récents logiciels de statistique. Cet article donne un aperçu de l'ANCOVA classique et présente des applications plus complexes de cette dernière dans des modèles mixtes. Les auteurs décrivent trois applications évoluées, soit (i) le recours à l'ANCOVA pour détailler la réaction au dosage dans divers traitements, (ii) la stabilité du traitement dans de multiples environnements et (iii) la suppression de la variation spatiale que le blocage ne contrôle pas entièrement. Ces applications révèlent que l'ANCOVA soit est plus simple, soit donne plus de renseignements que les méthodes statistiques usuelles. Les auteurs présentent les principes et la théorie sous-jacents à l'analyse par modèle mixte de l'ANCOVA ainsi que des programmes SAS dans deux annexes afin de mieux illustrer l'usage de cette puissante technique en recherche agronomique et agricole, et en permettre une meilleure compréhension.

Mots clés: Analyse de la covariance, réaction au dosage, modèles mixtes, analyse du plus proche voisin, polynômes orthogonaux, variabilité spatiale, analyse de la stabilité, contrôle statistique des erreurs

In agronomy and crop research, precise experimentation is required to detect true treatment effects and true differences between treatments. Given the sample error variance (s^2) and the number of replicates (n), the precision of an experiment is measured by the standard error of a single treatment mean, $\sqrt{s^2/n}$, or by the standard error of the difference between a pair of treatments, $\sqrt{2s^2/n}$. In other words, the precision of an experiment can be increased by increasing the number of replicates or reducing the error variance. The error reduction is achieved by (i) careful selection of treatments; (ii) fine-tuning of experimental techniques; (iii) careful selection of experimental materials; (iv) careful selection of experimental units; (v) planned grouping of homogeneous experimental units; and (vi) taking additional observations. The first five methods for controlling the error variance are usually described and discussed in textbooks on experimental designs (e.g., Cochran and Cox 1957; Petersen 1994; Williams et al. 2006). The last technique is to reduce the error variance by removing part of the variability in the original variable (Y) associated with one or more independent variables Xs (additional observations). The technique is

Abbreviations: **ANCOVA**, analysis of covariance; **ANOVA**, analysis of variance; **NNA**, nearest neighbour analysis; **RCBD**, randomized complete block design; **SE**, standard error; **SS**, sum of squares

doi:10.4141/CJPS2010-032

known as the analysis of covariance (ANCOVA), the topic of this paper.

Analysis of covariance is a statistical technique that combines the methods of the analysis of variance (ANOVA) and regression analysis, and is sometimes called "ANOVA with covariates". This analysis typically includes two types of independent variables: (i) class or dummy variables whose levels are used to identify different treatments; and (ii) continuous variables, called covariates, that are directly measured. If all independent variables are dummy variables, then ANCOVA becomes ANOVA. If all independent variables are covariates, then ANCOVA becomes regression analysis. Like ANOVA, ANCOVA was invented by British statistician R. A. Fisher in the early 1930s and was described in his classic book "Statistical Methods for Research Workers" (Fisher 1932, §49.1). Since then, the technique has been further developed and expanded for application in agriculture and other disciplines (for reviews see Cochran 1957; Cox and McCullagh 1982; Milliken and Johnson 2002). Typical applications of ANCOVA include (i) the removal of extraneous variation that cannot be effectively controlled by experimental design, (ii) adjustment of treatment means by a common covariate value for equitable comparisons, (iii) testing for homogeneity of covariate or slopes for different treatments or treatment combinations, and (iv) the estimation of missing values. The last application has become less useful now that most present-day statistical software can easily handle unbalanced data.

While most statistics textbooks (e.g., Snedecor and Cochran 1980; Gomez and Gomez 1984; Steel et al. 1997) have one chapter that is exclusively devoted to ANCOVA, there are books (e.g., Milliken and Johnson 2002) that are totally devoted to the subject. Nevertheless, ANCOVA is an advanced topic that is only taught cursorily or ignored completely in many statistics classes. Additionally, its application to agronomy and crop research is rarely described in these books. Several statistical software packages have been developed to greatly facilitate the use of ANCOVA for the analysis of research experiments. In particular, the recent development of the MIXED procedure in the SAS® system or similar procedures in other software packages (e.g., ASREML, GENSTAT, R, S-PLUS and SPSS) has had a great influence on how statistical analyses are performed as mixed models provide a common framework for many analyses of designed and observational experiments (Littell et al. 2006; Yang 2010). Unfortunately popular textbooks such as Steel et al. (1997) provide little coverage on the use of ANCOVA under the mixed-model framework. As a result, many agronomists and crop scientists are not aware of capabilities available in statistical software for implementation of ANCOVA or even if the analysis is done, they do not provide adequate or valid interpretation of the output.

The objectives of this paper are to provide an overview of conventional ANCOVA and to introduce more advanced uses of ANCOVA under mixed models. Three applications have been selected from agronomy and crop research to illustrate why the advanced analyses are more appropriate or advantageous. We will illustrate the mixed-model analysis using the SAS system (SAS Institute, Inc. 2008) because this software has been widely used by agronomists and other crop researchers. A technical appendix (Appendix A) is provided for those who wish to have some in-depth understanding of principles and theory underlying mixed-model analysis of ANCOVA. All the SAS programs for implementation of ANCOVA for the analysis of the four examples discussed in the paper are listed in Appendix B. Interested researchers or data analysts can adopt or modify these programs to accommodate their specific analytical needs.

CONVENTIONAL ANALYSIS

Statistical Models

Let us start with an example from Snedecor and Cochran (1980), Table 18.5.2) of a one-way classification experiment in a randomized complete block design (RCBD). The experiment compared the yields Y (lbs/plot) of six varieties of corn (*Zea mays* L.). There was some plot-to-plot variation in number of plants (X). Suppose that higher plant density leads to higher yield per plot under higher soil fertility and there were soil fertility differences between plots. Precision would be improved by adjusting yield with the covariance of plant number where plant number is a surrogate for the fertility level of each plot. A mathematical model to describe these data is given by:

$$Y_{ij} = \mu + \alpha_i + b_j + \beta_i(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij} \tag{1}$$

where Y_{ii} is the yield of the ith variety in the jth block and X_{ii} is the plant number (auxiliary variate or covariate) in the same plot, μ is the overall mean yield, α_i is the effect of the ith variety, b_j is the effect of the jth block, β_i is the slope of the ith variety, $\bar{X}_{..}$ is the overall mean over all X_{ij}s, and ε_{ij} is the experimental error. If the plant numbers (X_{ij}s) were not measured, then the variation of Y due to $\beta_i(X_{ij} - \bar{X}_{..})$could not be determined and would be included in the error term. In this case, model (1) becomes an ANOVA model

$$Y_{ij} = \mu + \alpha_i + b_j + e_{ij}. \tag{2}$$

Since X and Y are known to be closely related, model (1) would fit the Y values better than model (2), and the residuals ε_{ij} should be smaller than e_{ij}. If all the varieties have the same response (i.e., $\beta_1 = \beta_2 = \ldots = \beta_6 = \beta$), then model (1) is simplified to become

$$Y_{ij} = \mu + \alpha_i + b_j + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}. \tag{3}$$

Brief Review of Principles and Procedures

Model (3) is the baseline model for the classical ANCOVA analysis (Cochran and Cox 1957; Snedecor and Cochran 1980, Ch. 18; Steel et al. 1997, Ch. 17). The first step is to estimate the regression coefficient (β). In the usual regression analysis, the regression coefficient in a single sample is estimated as $\hat{\beta} = \Sigma\Sigma_{ij}x_{ij}y_{ij}/\Sigma\Sigma_{ij}x_{ij}^2$, where $x_{ij} = (X_{ij} - \bar{X}_{..})$ and $y_{ij} = (Y_{ij} - \bar{Y}_{..})$. The reduction in the sum of squares of Y due to the regression is $(\Sigma\Sigma_{ij}x_{ij}y_{ij})^2/\Sigma\Sigma_{ij}x_{ij}^2$. These estimates remain true for multiple samples or multiple classifications in ANCOVA, but β is now estimated using the error sum of products of X and Y (E_{xy}) and error sum of squares of X (E_{xx}), $\hat{\beta} = E_{xy}/E_{xx}$. The reason for this use of error components for the estimation of β is that with a covariance adjustment the values of $Y - \beta x$, instead of Y, are used to estimate and compare treatment effects. Thus ANCOVA is the same as an ANOVA for the quantity $Y - \beta x$. The value of β can be chosen to minimize the error sum of squares of the new variable $Y - \beta x$,

$$\begin{aligned} E_{y|x} &= E_{yy} - 2\beta E_{xy} + \beta^2 E_{xx} \\ &= E_{yy} - \frac{E_{xy}^2}{E_{xx}} + E_{xx}\left[\beta^2 - 2\beta\frac{E_{xy}}{E_{xx}} + \left(\frac{E_{xy}}{E_{xx}}\right)^2\right] \\ &= E_{yy} - \frac{E_{xy}^2}{E_{xx}} + E_{xx}\left(\beta - \frac{E_{xy}}{E_{xx}}\right)^2 \end{aligned} \quad (4)$$

The least square estimate of β is $\hat{\beta} = E_{xy}/E_{xx}$ and the minimum value of $E_{y|x}$ is $E_{y|x} = E_{yy} - E_{xy}^2/E_{xx}$. Thus the conventional version of ANCOVA is best described using the format of bivariate analysis of variance of X and Y. The necessary formulas and computations are given in Table 1 and the SAS statements implementing these ANCOVA calculations are given in Appendix B.1.

Error Control

The results from the SAS analysis of corn data are summarized at the bottom of Table 1. As significant variation among the plant numbers (X) within varieties may distort the yields (Y), we must first look at the F value for varieties with X. The mean squares are 45.83/5 = 9.17 for varieties and 113.83/15 = 7.59 for error, giving an F value of 9.17/7.59 = 1.21. This F value is not significant, indicating that the variations in plant numbers were largely random and that the covariance adjustment for plant number will not cause bias.

A key part of ANCOVA is the use of the variety + error line for computing the F value of the adjusted means (Table 1). The error sum of squares adjusted for regression is: $E_{y|x} = E_{yy} - E_{xy}^2/E_{xx} = 8752.3 - 917.3^2/113.8 = 1361.3$. The variety + error sum of squares adjusted for regression is: $(T+E)_{y|x} = (T_{yy} + E_{yy}) - (T_{xy} + E_{xy})^2/(T_{xx} + E_{xx}) = (9490.0 + 8752.3) - (559.3 + 917.3)^2/(45.8 + 113.8) = 4588.5$. The F test for adjusted means is: $F_{y|x} = [((T+E)_{y|x} - E_{y|x})/(t-1)]/[E_{y|x}/((r-1)(t-1)-1)] = [(4588.5 - 1361.3)/5]/(1361.3/14) = 6.64$. The regression adjustment leads to a considerable reduction in the error mean square: $E_{yy}/[(r-1)(t-1)] = 8752.3/15 = 583.5$ to $E_{y|x}/[(r-1)(t-1)-1] = 1361.3/14 = 97.2$ with a corresponding increase in $F_{(y|x)} = 6.64$ from $F_{(yy)} = 3.25$.

Adjusted Treatment Means

The adjusted mean for the ith treatment is calculated using the following formula

$$\bar{Y}_{i\cdot(\text{adj.})} = \bar{Y}_{i\cdot} - \hat{\beta}(\bar{X}_{i\cdot} - \bar{X}_{..}). \quad (5)$$

This adjusted mean has a standard error (SE_i)

$$SE_i = \sqrt{\frac{E_{y|x}}{df_e - 1}\left[\frac{1}{r_i} + \frac{(\bar{X}_{i\cdot} - \bar{X}_{..})^2}{E_{xx}}\right]}, \quad (6)$$

where r_i is the number of replications for the ith treatment and df_e is the error degrees of freedom.

Table 1. Bivariate analysis of variance (unadjusted analysis) with SS_x, SS_y and CP_{xy} being sums of squares for X and Y and cross products between X and Y, and adjusted analysis with $SS_{y|x}$, and $MS_{y|x}$ being the sum of squares and mean squares of Y adjusted for X for the plant density data given in Table 18.5.2 of Snedecor and Cochran (1980). Data analyses using PROC GLM of SAS with the programming described in Appendix B.1

	Unadjusted analysis				Adjusted analysis				
Source	df	SS_x	CP_{xy}	SS_y	df	$SS^z_{y\|x}$	$MS_{y\|x}$	F	P value
General formula									
Block	(r − 1)	B_{xx}	B_{xy}	B_{yy}					
Variety (T)	(t − 1)	T_{xx}	T_{xy}	T_{yy}	(t − 1)	$T_{y\|x}$	$MST_{y\|x}$		
Error (E)	(r − 1)(t − 1)	E_{xx}	E_{xy}	E_{yy}	(r − 1)(t − 1) − 1	$E_{y\|x}$	$MSE_{y\|x}$		
T + E	r(t − 1)	$T_{xx}+E_{xx}$	$T_{xy}+E_{xy}$	$T_{yy}+E_{yy}$	r(t − 1) − 1	$(T+E)_{y\|x}$	$MS(T+E)_{y\|x}$		
Regression					1	$R_{y\|x}$	$MSR_{y\|x}$		
For the data set									
Block	3	21.7	8.5	436.2					
Variety (T)	5	45.8	559.3	9490.0	5	3227.2	645.5	6.64	0.0023
Error (E)	15	113.8	917.3	8752.3	14	1361.3	97.2		
T + E	20	159.7	1476.5	18242.3	19	4588.5			
Regression					1	7391.1	7391.1	76.02	4.96×10^{-7}

[z] $E_{y|x} = E_{yy} - (E_{xy})^2/E_{xx}$, $(T+E)_{y|x} = (T_{yy}+E_{yy}) - (T_{xy}+E_{xy})^2/(T_{xx}+E_{xx})$, $R_{y|x} = E_{yy} - E_{y|x}$, and $T_{y|x} = (T+E)_{y|x} - E_{y|x}$.

The difference between the adjusted means for the *i*th and *j*th treatments is given by

$$\bar{\bar{Y}}_{i\cdot(\text{adj.})} - \bar{\bar{Y}}_{j\cdot(\text{adj.})} = \bar{\bar{Y}}_{i\cdot} - \bar{\bar{Y}}_{j\cdot} - \hat{\beta}(\bar{\bar{X}}_{i\cdot} - \bar{\bar{X}}_{j\cdot}). \tag{7}$$

This difference has a standard error (SE_{ij})

$$SE_{ij} = \sqrt{\frac{E_{y|x}}{df_e - 1}\left[\frac{1}{r_i} + \frac{1}{r_j} + \frac{(\bar{\bar{X}}_{i\cdot} - \bar{\bar{X}}_{j\cdot})^2}{E_{xx}}\right]}. \tag{8}$$

In the example, the average yields before the adjustment for varieties A, B, C, D, E and F, were 173.0, 182.3, 194.5, 232.8, 201.0 and 215.0 lbs, respectively; the average plant numbers for the respective varieties were 24.0, 25.3, 26.5, 28.0, 27.8 and 26.5 with the overall average being 26.3 plants per plot. Thus, the adjusted mean for variety A is $173.0 - (8.06)(24.0 - 26.3) = 191.8$ with SE_A being $[97.2(1/4 + (-2.3)^2/113.8)]^{0.5} = 5.38$. The difference between the adjusted means for varieties A and B is $191.8 - 191.0 = 0.8$ with SE_{AB} being $[97.2 (2/4 + (24.0 - 25.3)^2/113.8)]^{0.5} = 7.07$. All pairwise comparisons (Table 2) indicated that the two highest yielding varieties D and F did not differ significantly but they were significantly better than the remaining four varieties, which did not differ significantly among themselves. For easy calculation of average SE for comparing the differences between any pair of adjusted means, Snedecor and Cochran (1980), p. 377) used an approximate formula calculated as $[(2/r)MSE_{y|x}(1 + (T_{xx}/(t-1))/E_{xx})]^{0.5} = [(2/4)(97.2)(1 + 45.8/5/113.8)]^{0.5} = 7.25$. This average SE is too high for some pairs of adjusted means but too low for other pairs when compared with the adjusted pairwise comparisons of Table 2.

Relative Efficiency of ANCOVA vs. ANOVA

The effectiveness of ANCOVA as a means of error control relative to ANOVA can be made by comparing the variance of a treatment mean with and without the covariance adjustment. For the corn yield data (Table 1), the error mean square without the covariance adjustment was 583.49 whereas the error mean square with the covariance adjustment was 97.2. This last value must be readjusted upward to allow for sampling variability in the regression coefficient [see Snedecor and Cochran (1980), p. 369 for an explanation]. The readjustment entails using the treatment and error sums of squares from the ANOVA for X ($T_{xx} = 45.83$ and $E_{xx} = 113.83$) (Table 1). Thus, the effective error mean square after covariance adjustment is $MS_E = MSE_{y|x}[1 + T_{xx}/(t-1)/E_{xx}] = 97.2{*}[1 + 45.8/5/113.83] = 105.06$. The relative efficiency of ANCOVA compared with ANOVA is then $[E_{yy}/(r-1)(t-1)]/MS_E{*}100\% = (583.49/105.06){*}100\% = 555\%$. In other words, ANCOVA with four replications gives as precise an estimate as unadjusted means with 22–23 replications.

ADVANCED ANALYSIS

ANCOVA can be carried out using PROC GLM or PROC MIXED of the SAS® system (SAS Institute, Inc. 2008). However, in some popular textbooks (e.g., Steel et al. 1997), the use of PROC GLM for ANCOVA is described in detail but with little or no mention of PROC MIXED. In agronomy and crop research, ANCOVA can be applied to the analysis of experimental designs with fixed treatment effects and some kind of blocking. If blocking is used, then more than one size of experimental unit exists. In the corn example, the experimental design is a RCBD; thus, blocks and plots within the blocks are two sizes of experimental unit. As blocking is often considered as random, a mixed-effects model is more appropriate for the analysis. Furthermore, in the mixed model we are able to construct the estimate of the slope by combining the intra-block information with the inter-block information (Littell et al. 2006, p. 258–263). In Appendix A, we provide a description of how a linear model for ANCOVA can be built from an actual data set (corn example) and show the differences between models used by PROC GLM and PROC MIXED. This should facilitate an in-depth understanding of more elaborate applications of ANCOVA. In the following we will describe three such applications.

Application # 1: Analysis of Dosage Response

Orthogonal Polynomial Analysis

Gomez and Gomez (1984, p. 317–327) described an evaluation experiment where the effect of five nitrogen (N) rates (0, 60, 90, 120 and 150 kg ha^{-1}) on rice (*Oryza sativa* L.) yield (tone ha^{-1}) was studied in two seasons, one dry and one wet. Each trial had a randomized

Table 2. Adjusted mean yields (lbs/plot) of six corn varieties on the diagonal (±standard error), differences between pairs of the six adjusted means (below the diagonal) and probabilities that pairwise differences are zero (above the diagonal). Data analyses using PROC MIXED of SAS programming as given in Appendix B.1

Variety	A	B	C	D	E	F
A	191.80±5.38	0.9090	0.8562	0.0036	0.7800	0.0100
B	0.82±7.07	190.98±5.03	0.7625	0.0019	0.8521	0.0063
C	−1.36±7.35	–2.18±7.07	193.16±4.93	0.0025	0.6211	0.0108
D	−27.52±7.89	−28.34±7.42	−26.16±7.11	219.32±5.17	0.0008	0.4390
E	2.22±7.79	1.39±7.35	3.57±7.07	29.74±6.98	189.58±5.10	0.0043
F	−21.86±7.35	−22.68±7.07	−20.50±6.97	5.66±7.11	−24.07±7.07	213.66±4.93

complete block design (RCBD) with three replications. A combined analysis over the two planting seasons was carried out (i) to examine if there was differential yield response to N rates in the dry and wet seasons and (ii) to determine if it was necessary to have separate N recommendations for the two seasons.

Yields were plotted against N rates for individual replicates and planting seasons (Fig. 1). Since there are five N levels, a yield response can be described using a polynomial up to the fourth degree. Gomez and Gomez (1984) partitioned the sum of squares (SS) for N rates using orthogonal contrasts due to linear, quadratic and higher-order regression effects: Total N SS=SS (linear)+SS (quadratic)+SS (remainder). Most statistical textbooks give a table of orthogonal polynomial coefficients for balanced data with equally spaced treatment levels. However, using conventional regression analyses, orthogonal polynomial coefficients for unbalanced data with unequally spaced treatment levels (as in this case) are difficult to derive. Gomez and Gomez (1984, p. 229–233) described a lengthy procedure of how to obtain orthogonal polynomial coefficients for their four unequally spaced levels, but, as these authors admitted, the contrast coefficients for other unequally spaced levels must be re-derived. As well, it is desirable to estimate regression equations as part of a combined analysis rather than using separate regression analysis as done by Gomez and Gomez (1984).

For a SAS user who is still interested in orthogonal polynomial analysis, a SAS IML function called ORPOL (SAS Institute, Inc. 2008) can be used to obtain orthogonal polynomial coefficients for unequally spaced treatment levels. For our current example with five unequally spaced N levels, the SAS code is as follows:

```
proc iml; levels = { 0,60,90,120,150} ; coef = orpol
(levels`); print coef; quit;
```

The outputs consist of five columns. The first column is a normalized polynomial of degree 0 (a constant polynomial) evaluated at each N rate; the second column is a normalized polynomial of degree 1 (linear) evaluated at each N rate; the third column is a normalized polynomial of degree 2 (quadratic) evaluated at each N rate; and so on. As the zeroth-degree polynomial is used to calculate the mean value, we only present coefficients for the first- to fourth-degree normalized orthogonal polynomials (Table 3). These normalized sets are equivalent to the non-normalized ones of Gomez and Gomez (1984, see p. 323). For example, the non-normalized coefficients for the linear response (−14 −4 1 6 11) can be converted into the normalized set by dividing each value by a normalization factor of $370=[(-14)^2+(-4)^2+1^2+6^2+11^2]$. Three properties about a set of orthogonal polynomial coefficients can be verified numerically using the coefficients in Table 3: (i) the coefficients for each contrast sum to zero, for example, $(-0.7278)+(-0.2080)+0.0520+0.3119+0.5719=0$; (ii) the coefficients are orthonormal because the squared coefficients for each contrast sum to one, for example, $(-0.7278)^2+(-0.2080)^2+(0.0520)^2+(0.3119)^2+(0.5719)^2=1$; and (iii) the orthogonality means that the sum of cross-products of coefficients in any pair of rows in Table 3 is zero, for example, the sum of cross-products of linear and quadratic coefficients is $(-0.7278{*}0.4907)+(-0.2080{*}-0.4729)+(0.0520{*}-0.4595)+(0.3119{*}-0.1160(+(0.5719{*}0.5576)=0$.

The CONTRAST statement in PROC GLM or PROC MIXED can be used to partition the total SS for N rates into components due to linear, quadratic,

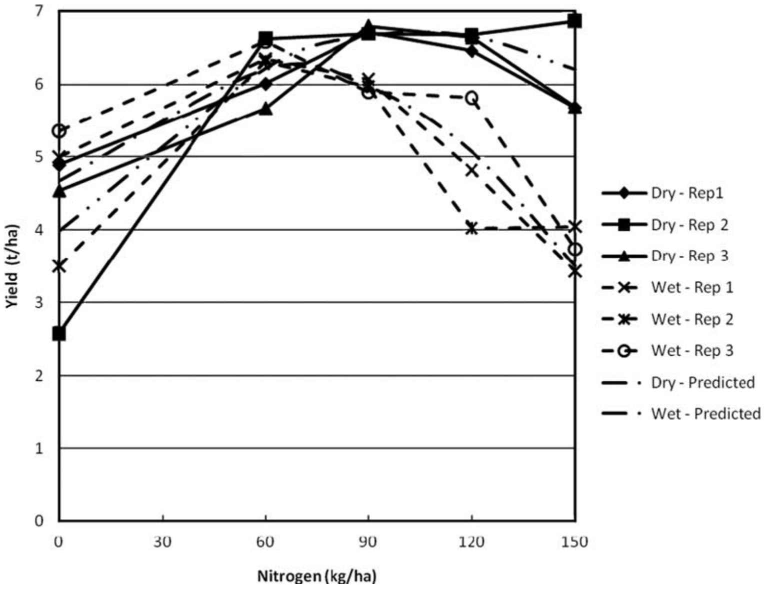

Fig. 1. Observed and predicted yield of rice tested with five rates of nitrogen in dry and wet crop seasons.

Table 3. Orthogonal polynomial coefficients and their properties for five nitrogen levels. Contrasts determined by PROC ILM of SAS as shown in Appendix B.2.1. Orthonormals determined by squaring the Contrast coefficient, Orthogonality computed by multiplying contrast coefficients

Nitrogen level	0	60	90	120	150	Total
(i) Contrast						
Linear (L)	−0.7278	−0.2080	0.0520	0.3119	0.5719	0
Quadratic (Q)	0.4907	−0.4729	−0.4595	−0.1160	0.5576	0
Cubic (C)	−0.1677	0.6312	−0.2170	−0.6213	0.3748	0
Quartic (D)	0.0367	−0.3671	0.7342	−0.5507	0.1468	0
(ii) Orthonormal						
L^2	0.5297	0.0433	0.0027	0.0973	0.3271	1
Q^2	0.2408	0.2236	0.2111	0.0135	0.3109	1
C^2	0.0281	0.3984	0.0471	0.3860	0.1405	1
D^2	0.0013	0.1348	0.5390	0.3033	0.0216	1
(iii) Orthogonality						
L × Q	−0.3571	0.0984	−0.0239	−0.0362	0.3189	0
L × Q	0.1221	−0.1313	−0.0113	−0.1938	0.2143	0
L × Q	−0.0267	0.0764	0.0382	−0.1718	0.0840	0
Q × C	−0.0823	−0.2985	0.0997	0.0721	0.2090	0
Q × D	0.0180	0.1736	−0.3374	0.0639	0.0819	0
C × D	−0.0062	−0.2317	−0.1593	0.3421	0.0550	0

cubic and quartic responses based on the orthogonal polynomial coefficients given in Table 3. Such partitioning is made possible due to the orthogonality among the different contrasts. In specifying the coefficients for each CONTRAST statement, the sum of coefficients for each contrast must be numerically zero. Otherwise PROC GLM or PROC MIXED will declare the contrast as non-estimable and will not provide an output. If a non-zero sum occurs, simply change one coefficient value so that the sum is exactly zero! This tiny alteration will have a negligible effect on the SS and *F* values.

ANCOVA Approach

As shown above, the use of orthogonal polynomial coefficients is able to evaluate the linear, quadratic and higher-order responses to the N rates. However, this may not be the most effective means for the evaluation. First of all, some efforts [a lengthy procedure of Gomez and Gomez (1984) or use of SAS IML functions] are needed to find the orthogonal polynomial coefficients for the unequally spaced treatment levels. Second, when a significant response is confirmed, the regression equation needs to be estimated for prediction. The ANCOVA approach can be used to partition the total treatment SS into different response components and to estimate the regression equation. Finally, and perhaps most importantly, when the data are unbalanced (e.g., missing values due to biotic or abiotic damage or any unforeseen loss), the orthogonal polynomial analysis may not work because the imbalance would cause different orthogonal contrasts to be nonestimable. On the other hand, the ANCOVA approach always works regardless of whether or not the data are balanced.

The first and key step of the ANCOVA approach is to create a new variable (which we will call *n*) whose entries are identical to those of variable NITROGEN. The ANCOVA approach takes the advantage of the fact that variable NITROGEN is a quantitative variable and that it can serve as a classification variable or a covariate. However, the same variable cannot be a classification variable and a covariate simultaneously and thus a second and duplicated variable n is needed to serve as a covariate, leaving NITROGEN as a classification variable. The appropriate SAS code for the analysis of the nitrogen level example is given in Appendix B.2.1 and the core part of the code is as follows:

```
proc mixed method = type1;
class season nitrogen rep;
model y = season n n* n nitrogen season* n
season* n* n
     season* nitrogen;
```

The option of METHOD = TYPE1 in the PROC MIXED statement is needed to ensure the correct ANCOVA analysis. As it is unnecessary to test the significance of a polynomial term adjusted for a higher-order term, sequential sums of squares (Type I SS) are the natural choice and partial sums of squares (Type III SS) should be ignored in the polynomial response analysis (Steel et al. 1997, p. 387). Since variable n is not listed in the CLASS statement, the terms n and n*n in the MODEL statement are considered as direct regression variables or covariates, representing linear and quadratic response to N rates. As the variable NITROGEN is a CLASS variable it serves as a ‘‘lack of fit’’ variable to capture the remaining variation after fitting linear and quadratic responses. The interaction terms of season*n and season*n*n are used to evaluate heterogeneous linear and quadratic responses to N rates.

From the two analyses for covariance for yield of rice shown in Table 4, two statements can be made: (i) ANCOVA provides the same partitioning of the total treatment SS into linear, quadratic and higher-order

Table 4. Two analyses of covariance for yield response of rice to five rates of nitrogen in two growing seasons (dry and wet) (from Gomez and Gomez 1984). Analyses I obtained from running the orthogonal polynomial analysis of PROC GLM of SAS as outlined in Appendix B.2.1. Analyses II obtained from running the ANCOVA approach to dosage response analysis of PROC GLM and PROC MIXED as outlined in Appendix B.2.1

Source	DF	Type I SS	Mean Square	*F*	*P* value
Analysis I					
Season	1	4.50	4.50	14.26	0.0195
Rep(season)	4	1.26	0.32	–	–
Nitrogen(n)	4	18.75	4.69	10.62	0.0002
n	1	1.44	1.44	3.25	0.0901
N*n	1	17.17	17.17	38.90	0.0000
Remainder	2	0.14	0.07	0.16	0.8548
Nitrogen*season	4	9.66	2.41	5.47	0.0057
n*season	1	8.81	8.81	19.95	0.0004
n*n*season	1	0.55	0.55	1.25	0.2810
Remainder*season	2	0.30	0.15	0.34	0.7190
Pooled error	16	7.06	0.44	–	–
Analysis II					
Season	1	4.50	4.50	14.26	0.0195
Rep(season)	4	1.26	0.32	–	–
Nitrogen	4	18.75	4.69	10.62	0.0002
n	1	1.44	1.44	1.49	0.2897
n*n	1	17.17	17.17	145.37	0.0003
Remainder	2	0.14	0.07	0.21	0.8186
Nitrogen*season	4	9.66	2.41	5.47	0.0057
n*season	1	8.81	8.81	9.12	0.0392
n*n*season	1	0.55	0.55	4.65	0.0972
Remainder*season	2	0.30	0.15	0.44	0.6610
Pooled error	16	7.06	0.44	–	–
n*rep(season)	4	3.86	0.97	–	–
n*n*rep(season)	4	0.47	0.12	–	–
Remainder*rep(season)	8	2.73	0.34	–	–

components as in the orthogonal polynomial analysis; and (ii) the significant season*nitrogen interaction effect is due to heterogeneity of linear responses between the two seasons. This latter finding supports the need for an estimation of separate regression equations for wet and dry seasons. The following SAS statements can now be run to obtain the required estimates for intercepts and regression coefficients:

```
proc mixed data = new method = type1;
class season nitrogen rep;
model  y = season  n(season)  n*n(season)  /noint
solution;
random rep(season);
```

In PROC MIXED, the RANDOM rep(season) statement and the NOINT option cause the intercepts to be estimated directly, rather than requiring ESTIMATE statements as in PROC GLM (Littell et al. 2002, p. 262). In addition, the RANDOM statement of PROC MIXED also produces valid SEs of the estimated regression coefficients if blocks are considered random. In contrast, PROC GLM always computes SEs based on the assumption of fixed blocks and would underestimate the SEs of random blocks. The SOLUTION option under the MODEL statement outputs estimated regression coefficients for obtaining quadratic regression equations for both dry and wet seasons, that are the same as those given by Gomez and Gomez (1984, p. 324):

$$\text{Dry: } Y = 3.9825 + 0.05233N - 0.00025N^2$$

$$\text{Wet: } Y = 4.6749 + 0.04772N - 0.00037N^2$$

The predicted yields from these two equations are overlayed with observed values in Fig. 1 and show that the rate of yield increase with an increase in N rate is higher in the dry season than in the wet season. The nitrogen rate at which maximum yields occur (N_{max}) is obtained by differentiating the quadratic equation, $dY = d(c + bN + aN^2)$, or $dY/dN = b + 2aN$, setting the differential to zero ($dY/dN = b + 2aN = 0$) and solving for N, $N_{max} = -b/2a$. Thus, for the dry season, $N_{max} = -0.05233/2(-0.00025) = 102.5$ kg N ha^{-1} and for the wet season, $N_{max} = -0.04772/2(-0.00037) = 65.2$ kg N ha^{-1}. These results along with the significant n*season interaction support the need for different nitrogen recommendations for dry vs. wet seasons to optimize yield of rice at the location of the study.

To test if the interaction terms, n*season and n*n*season, are significant, we used a pooled error variance. However, if this pooled error variance is heterogeneous across different sources, then it needs to be further partitioned into components corresponding to n*season and n*n*season. This can be achieved by adding two terms, n*rep(season) and n*n*rep(season) in the RANDOM statement:

```
random rep(season) n* rep(season) n* n* rep
(season);
```

Thus, n*rep(season) is used as an error term for testing the significance of the *F*-test for the n*season effect and similarly, n*n*rep(season) will be used to test the n*n*season effect. The results from this further partitioning of the pooled error variance and corresponding tests are given in Analysis II of Table 4. The *F*-statistics and *P* values differ between Analysis I and Analysis II, but these differences are not large enough to alter the conclusion that different N-rate recommendations are needed for different environments.

What have we learned from Application #1?

ANOCVA is a much easier approach to studying dosage responses than conventional methods (e.g., orthogonal polynomial coefficient analysis). It avoids the need for constructing and implementing orthogonal contrasts representing linear, quadratic and higher-order responses to N rates. It works regardless of whether or not the data are balanced. The estimation of regression equations for prediction is an integral part of the ANCOVA approach and it is not a separate analysis as in the conventional analysis. The use of SAS PROC MIXED is recommended as it ensures that all tests are correct. It needs to be stressed that the METHOD = TYPE1 option (i.e., SS1) should be used for exact partitioning of the dosage SS into components due to linear, quadratic, etc. responses; the use of METHOD = TYPE3 (i.e., SS3) or REML option would produce nonsense results because it does not make sense for the linear effect to be adjusted for quadratic or higher-order terms!

Application #2: Stability of Treatments Across Environments

Why Stability Analysis?

In agronomy and crop research, many experiments are carried out using a complete or incomplete block design with a few replications at multiple sites and over several years. These multi-environmental experiments are needed to make recommendations about the treatment performance for future years over a wide region. The statistical analysis of such multi-environmental trials has received a great deal of attention, particularly in the analysis of genotype-by-environment interactions in plant breeding literature (e.g., Lin and Binns 1994; Smith et al. 2005). Plant breeders select genotypes that perform well across growing conditions. This selection can be achieved in two ways. The first and ideal approach is to divide the growing region (e.g., the Canadian prairies) into a number of agroclimatic zones within which adaphic (environmental) factors such as soil properties and weather conditions are relatively homogeneous. However, this approach has proven difficult to use because weather conditions may change greatly from year to year despite some stable adaphic properties. Therefore, most plant breeders take the second and more feasible approach by which genotypes with adaptation to a wide range of environmental conditions are selected. Generally, only those genotypes exhibiting stability over environments are retained during the selection process. We believe that agronomists and other crop scientists should also take the second approach of measuring stability of treatments across environments when developing recommendations on new agronomic practices or technologies. Here we will only describe one type of stability analysis pioneered by Yates and Cochran (1938) and popularized by Finlay and Wilkinson (1963) and Eberhart and Russell (1966). This stability analysis is also a nice application of the ANCOVA technique.

ANCOVA Approach to Stability Analysis

Littell et al. (2002, p. 420–431) described an analysis of data taken from a study that was carried out to compare three treatments (TRT) conducted at eight locations (LOC). At each location, a RCBD design was used, but the number of blocks varied: three blocks in locations 1 to 4, six blocks in locations 5 and 6, and 12 blocks in locations 7 and 8. The preliminary analysis by Littell et al. (2002) indicated a significant TRT × LOC interaction. More specifically, such an interaction was due to location-specific performance. For example, treatment 1 tended to be favored in "poor" locations but treatment 3 tended to be favored in "good" locations. This relationship can be analyzed using the ANCOVA approach in which an index characterizing the mean response at each location serves as a covariate. The ANCOVA model is:

$$\begin{aligned} Y_{ijk} = \mu + LOC_i + B(LOC)_{ij} + TRT_k + \beta_k I_i \\ + (TRT^*LOC)_{ik} + e_{ijk} \end{aligned} \quad (9)$$

where Y_{ijk} is the measured response of the *j*th replication of the *k*th treatment in the *i*th location ($i = 1, 2, \ldots, 8$; $j = 1, 2, \ldots, r_i$; $k = 1, 2, 3$), μ is the overall mean, LOC_i is the effect of the *i*th location, $B(LOC)_{ij}$ is the effect of the *j*th replication within the *i*th location, TRT_k is the effect of the *k*th treatment, $\beta_k I_i$ is the linear regression on a location index, $(TRT^*LOC)_{ik}$ is the interaction effect of the *i*th location with the *k*th treatment after fitting the regression, and e_{ijk} is the random error.

The location index is often unknown, but the mean response over all observations on each and every location can serve as a surrogate. The calculation can be easily implemented in the SAS system (Appendix B.2.2) as adopted from Littell et al. (2002). First, PROC SORT and PROC MEANS are used to produce a new data set, ENV_INDX, for the environmental coefficient, INDEX. Thus, SAS stores the means of Y on a location basis in the dataset ENV_INDX. Then, the data set ENV_INDX, is merged with the original data by location. Second, the PROC MIXED program is used for a mixed-model analysis where treatment effects are fixed and location and replication effects are random. Third, the term TRT*INDEX is put into the MODEL statement to examine if there are similar responses for different treatments (i.e., slopes are parallel). Fourth, the NOINT and SOLUTION options are employed to allow easier interpretation of SAS outputs. Fifth, the CONTRAST or ESTIMATE statement is used to compute a test to determine if the treatment effects are equal (a CONTRAST statement was used in Appendix B.2.2).

To assess how the linear regression of Y on INDEX (covariate) contributes to the variation among locations, we ran PROC MIXED twice, one with INDEX and the other without INDEX as a covariate being included in the MODEL statement (cf. in Appendix B.2.2). Looking at the "Covariance Parameter Estimates" from the two runs in the SAS Output Window, the estimates of LOC and LOC*TRT variances were 0 and 0.83 with inclusion of INDEX but were 63.75 and 34.43 without inclusion of INDEX. Therefore, the linear regression accounts for most of the variation among locations. The "Solution for fixed Effects" values are interpreted as follows: TRT estimates $\mu + TRT_k$ and INDEX*TRT estimates β_k. These estimates allow for writing out the prediction equations for the three treatments:

$$\text{Treatment 1: } \hat{Y}_1 = 12.4035 + 0.6345 I_i$$

$$\text{Treatment 2: } \hat{Y}_2 = 17.0483 + 0.6232 I_i$$

$$\text{Treatment 3: } \hat{Y}_3 = 29.4519 + 1.7423 I_i$$

Thus, we can employ these equations to predict the treatment means at a given location by using the index value for that location. In Table 5, we summarized the SAS calculation of the treatment means for the "average" location (the mean of the INDEX covariate over all locations), the "poor" location (the minimum INDEX value) and the "good" location (the maximum INDEX value). For treatment 1 at the "average" location (i.e., $I_{Ave.} = 45.2$), the mean value is predicted as $\hat{Y}_1 = 12.4035 + (0.6345)*(45.2) = 41.4$, which is the same as the LS mean from the SAS output. The slope (TRT*INDEX) estimate is much larger for treatment 3, but the intercept (TRT) is much smaller, suggesting that treatment 3 performed worse in poor locations but better in good locations than treatments 1 and 2. This conclusion can be verified using PROC MIXED and running LSMEANS statements with the AT option and different INDEX values (i.e., $I_{poor} = 30.9$ for the "poor" environment and $I_{good} = 57.9$ for the "good" environment).

What have we learned from Application #2?

We have illustrated the use of ANCOVA for studying the stability of treatments across different environments. Specifically, ANCOVA can be used to partition the total variability of treatment × location interaction into two components, one due to the linear regression of performance (Y) on location index and the other due to the residual left after fitting the regression. If the estimated regression is significant, then we recommend that the focus should be on examining changes in treatment responses at the different location types, by comparing and contrasting performances of different treatments in "poor", "average" and "good" environments as defined by specific values of the location indices.

The stability analysis has been extensively used in plant breeding literature but it is largely ignored in agronomy and other crop research even though many field trials are carried out in multiple locations over several years just like variety trials. The analyses shown in Application #2 along with SAS code in Appendix B.2.2 can be used, directly or with slight modifications, depending on individual experimental designs, by those who are interested in exploring this powerful tool.

Application #3: Analysis of Spatial Variability

Why Spatial Analysis?

The next data set is taken from a field pea (*Pisum sativum* L.) variety trial described by Yang et al. (2004) who analyzed a total of 157 trials. In this trial, the experimental design was a RCBD with four replications

Table 5. Stability of three treatments (±standard error) assessed at poor, average and good locations as defined by location indices. The raw data is taken from Littell et al. (2002). Output provided by analyses as given in Appendix B.2.2

			Comparisons between treatments			
Location	Treatment	Mean ± SE	Pair	Difference ± SE	*t*-test	*P* value
Poor	1	32.01 ± 1.79	1 vs. 2	−4.30 ± 2.54	−1.69	0.0987
(Index = 30.9)	2	36.31 ± 1.79	1 vs. 3	7.63 ± 2.54	3.01	0.0048
	3	24.38 ± 1.79	2 vs. 3	11.92 ± 2.54	4.70	3.74×10^{-5}
Average	1	41.08 ± 0.85	1 vs. 2	−4.14 ± 1.20	−3.45	0.0027
(Index = 45.2)	2	45.22 ± 0.85	1 vs. 3	−8.22 ± 1.20	−6.85	1.55×10^{-6}
	3	49.30 ± 0.85	2 vs. 3	−4.08 ± 1.20	−3.40	0.0030
Good	1	49.14 ± 1.69	1 vs. 2	−3.99 ± 2.40	−1.67	0.1114
(Index = 57.9)	2	53.13 ± 1.69	1 vs. 3	−22.28 ± 2.40	−9.30	1.67×10^{-8}
	3	71.43 ± 1.69	2 vs. 3	−18.29 ± 2.40	−7.64	3.30×10^{-7}

(blocks) and 28 varieties in each of the four blocks; thus, there were a total of $4 \times 28 = 112$ plots at each location. A RCBD was used in the other 156 trials as well, but block sizes varied from trial to trial with a range of 12 to 32 varieties per block. The usual analysis of these RCBD experiments may be problematic. In RCBD, proper blocking can reduce error by maximizing the difference between blocks and maintaining the plot-to-plot homogeneity within blocks, but blocking is ineffective if heterogeneity between plots does not follow a definite pattern (e.g., spotty soil heterogeneity; unpredictable pest incidence after blocking). In addition, when block size is large [>8–12 plots per block], intra-block heterogeneity is inevitable (Stroup et al. 1994)! Thus, efficiency of the RCBD is often poor in agronomy trials involving a large number of treatments. An incomplete block design, such as a lattice or one of the more flexible α-designs (Williams et al. 2006) may be used to have smaller blocks but spatial heterogeneity may persist in small blocks as well. Evidently, such "design-based" control of the error variation may not be sufficient to remove all spatial variation in such large variety or agronomic field trials.

There are different "model-based" analyses that exploit the information on plot positions to estimate and correct for spatial variation within and among blocks [for reviews on these methods see Stroup et al. (1994) and Yang et al. (2004)]. Here we describe one such analysis, known as the nearest neighbour analysis (NNA). This analysis makes direct use of the ANCOVA technique. In NNA, plot performance is adjusted for spatial variability by using information from the immediate neighbouring plots (Wilkinson et al. 1983; Brownie et al. 1993). Such adjustment is effective if the correlation between residuals for two adjacent plots (in one-dimensional adjustment) or four adjacent plots (in two-dimensional adjustment) is higher than that for plots far apart, as is the case in many field trials.

ANCOVA Approach to Spatial Analysis

The ANCOVA model for NNA is given by Brownie et al. (1993) as:

$$Y_{ij} = \mu + \tau_{k(ij)} + \beta_i X_{ij} + \varepsilon_{ij}, \tag{10}$$

where Y_{ij} is the observed response in the jth plot within the ith block or plot ij, the term $\mu + \tau_{k(ij)}$ represents the mean performance of the kth variety k in plot ij, β_i is the trend effect representing systematic spatial variation in the ith block, X_{ij} is the covariate (to be explained below) and ε_{ij} is the random residual. The covariate for the spatial trend effect is calculated as the average of residuals of the two nearest or neighbouring plots, $X_{ij} = (e_{i,j-1} + e_{i,j+1})/2$ where $e_{ij} = Y_{ij} - \bar{Y}_{(ij)}$ with $\bar{Y}_{(ij)}$ being the variety mean in plot ij. For border plots at either end of a block with only a neighbouring plot on one side, the covariate is calculated as the average of the residuals for the two nearest plots on that one side. The SAS code for these calculations using PROC MIXED is given in Appendix B.2.3. The error variance from the NNA analysis of the field pea data (254,916) is about 37% smaller than that from the usual RCBD analysis (407,639), suggesting that a significant amount of intra-block spatial variability from this field trial was removed by NNA analysis.

Yang et al. (2004) calculated the efficiency of NNA over RCBD for all 157 field trials that were run across Alberta from 1997 to 2001. They measured the efficiency of the NNA analysis relative to the RCBD analysis by $1 - SSE_a/SSE_u$, where SSE_a and SSE_u are the adjusted (NNA) and unadjusted (RCBD) error sum of squares, respectively. According to Yang et al. (2004), the NNA analysis removed 28% of the residual variation due to spatial heterogeneity for the trials in 1997 and 33% in 1998, but only 12 to 13% for the trials in 1999 to 2001. The trials from 1997 and 1998 had much larger block sizes (28–32 varieties per block) than did those from the latter 3 yr (12–22 varieties per block) because green and yellow varieties were included in the same trials in the first 2 yr, but separated into different trials in the latter 3 yr. As expected, the averaged coefficients of variation of raw data were greater in 1997 and 1998 (15.9 and 17.7%) than in 1999 to 2001 (7.7 to 9.1%). Yang et al. (2004) concluded that while it is impossible to preclude the year-to-year variation, it appeared that the contrasting patterns of spatial variation between 1997 and 1998 and 1999 to 2001 were largely due to the differences in block sizes.

The adjusted (NNA) and unadjusted (RCBD) variety means along with their ranks are presented in Table 6. While Pearson correlation between adjusted (NNA) and unadjusted (RCBD) variety means is quite high (0.95) and Spearman correlation between their ranks is high as well (0.89), some rank changes do occur. For example, variety 5 is ranked as 11th highest before NNA adjustment but ranked down to 20th place after NNA adjustment.

What have we learned from Application #3?

NNA is an application of the ANCOVA technique using the information of neighbouring plots as a covariate in block designs. The NNA is able to account for plot-to-plot spatial variability within blocks, thereby further reducing experimental error. Many agronomy trials may also have a large number of treatments or treatment combinations (e.g., two or more production treatments in a factorial structure) to be included in a large block, thereby requiring a NNA or other spatial analyses for the removal of intra-block spatial heterogeneity. Treatment means may be ranked differently before and after adjustment for spatial variability.

CONCLUSION

The outcomes of agronomy experiments depend not only on the effects of the treatments, but also on extraneous variations (commonly known as experimental errors),

Table 6. Mean yields (kg ha^{-1}) of 28 field pea varieties in a randomized complete block design without adjustment (RCBD) and with nearest neighbour adjustment (NNA) for spatial variation along their respective ranks (Rank$_{RCBD}$ and Rank$_{NNA}$). Data from Yang et al. (2004)

Variety ID	RCBD	NNA	Rank$_{RCBD}$	Rank$_{NNA}$
1	6226	5796	2	5
2	5003	4948	21	24
3	5707	5478	8	12
4	6184	6017	3	2
5	5536	5195	11	20
6	5082	5100	20	21
7	5102	5392	19	15
8	5616	5654	10	7
9	5739	5901	6	4
10	6013	5972	4	3
11	4743	4832	26	25
12	5349	5394	15	14
13	4909	5030	24	23
14	5620	5697	9	6
15	5172	5260	18	18
16	5723	5539	7	11
17	5220	4799	17	26
18	5517	5567	12	9
19	6287	6053	1	1
20	4274	4593	27	27
21	5405	5540	13	10
22	5277	5268	16	17
23	4998	5401	22	13
24	2862	2863	28	28
25	5375	5390	14	16
26	4879	5038	25	22
27	5772	5600	5	8
28	4960	5229	23	19

which tend to mask the effects of the treatments. To minimize or control the experimental errors, field trials have generally used one or multiple forms of blocking including RCBD, split plots, or incomplete block designs (e.g., Cochran and Cox 1957; Petersen 1994; Williams et al. 2006). However, such design-based approaches may not always be effective in reducing the experimental error when blocking cannot cope with plot-to-plot variation within the blocks that are generated by competition between varieties, soil variation and fertility, or weather-related conditions such as rainfalls and snow cover. However, by using ANCOVA (model-based approach) the experimental error may be reduced. Blocking and ANCOVA can both be used to control the experimental error as they are complementary to each other. As blocking is decided prior to the start of an experiment, it can be used only to cope with sources of variation that are known or predictable. In contrast, ANCOVA can handle unpredictable sources of variation that occur during the course of the experiment so long as there is one covariate, or more, that represents the heterogeneity between plots and is quantitatively measurable. Thus, ANCOVA is a useful supplementary procedure to further reduce errors that cannot be removed by blocking.

ANCOVA can also be used for adjustment of treatment means. When a primary variable (Y) is known to be related with one or more covariates (Xs), treatment means in Y can be adjusted to a common X value, thereby producing fair and equitable comparisons among the treatments. While the use of adjusted means allows for the removal of bias arising from unequal covariate means among the treatments, adjusted means may not always be appropriate. Snedecor and Cochran (1980, p. 377) and Littell et al. (2002, p. 236) both warn that if the covariate means themselves depend on the treatments, adjustment is likely to be misleading. In our plant density example, yield is affected by the population density of the plants, but different varieties may have inherent (genetic) differences in their response to plant densities. Adjustment of mean yield to a common plant density that is outside the response range for some varieties would distort differences among varieties. In the corn example, however, adjusting yield is probably justified because (i) it is not known if the six varieties used differ significantly in response to plant density and (ii) the plot-to-plot variation in plant density is likely due to the fertility level of plots.

The three applications to agronomy and crop research described in this paper are only a small sample of many possible applications of ANCOVA. For example, Milliken and Johnson (2002) and Littell et al. (2006) have also described the use of ANCOVA for the analysis of more complex designs including split-plot, strip-plot, incomplete blocks and repeated measures in agriculture and other disciplines. Nevertheless, the underlying principles are the same for all applications. Our three selected case studies serve to illustrate the power of ANCOVA as a strategy for analyzing data from a variety of agronomy trials. In particular, we have focused on two major types of applications. First, a more familiar use of ANCOVA is for the removal of the extraneous variability in the experimental units that cannot be controlled or removed by a design structure such as blocking (e.g., the use of neighbouring plot information as a covariate for removing spatial variation in Application #3). Second, a less appreciated but more general application of ANCOVA is to determine and compare a series of regression models, one for each treatment or treatment combination (see the use of unequal slopes models in Applications #1 and #2). The strategy for the second application as adapted from Littell et al. (2006) is summarized in Fig. 2. The guidelines can be easily extended to the cases of more than one covariate.

In conclusion, ANCOVA is not only an effective means of improving the precision of experiments by removing extraneous variation that is not readily controlled by experimental design, but it is also a powerful statistical method for investigating dosage responses, stability of treatments across multiple environments, and spatial variation in agronomy trials. In comparison to conventional analysis, ANCOVA provides both simpler analyses in some applications and more information in others. We believe that ANCOVA is underutilized. It is hoped that this paper will help agronomists and other crop researchers to appreciate the power and

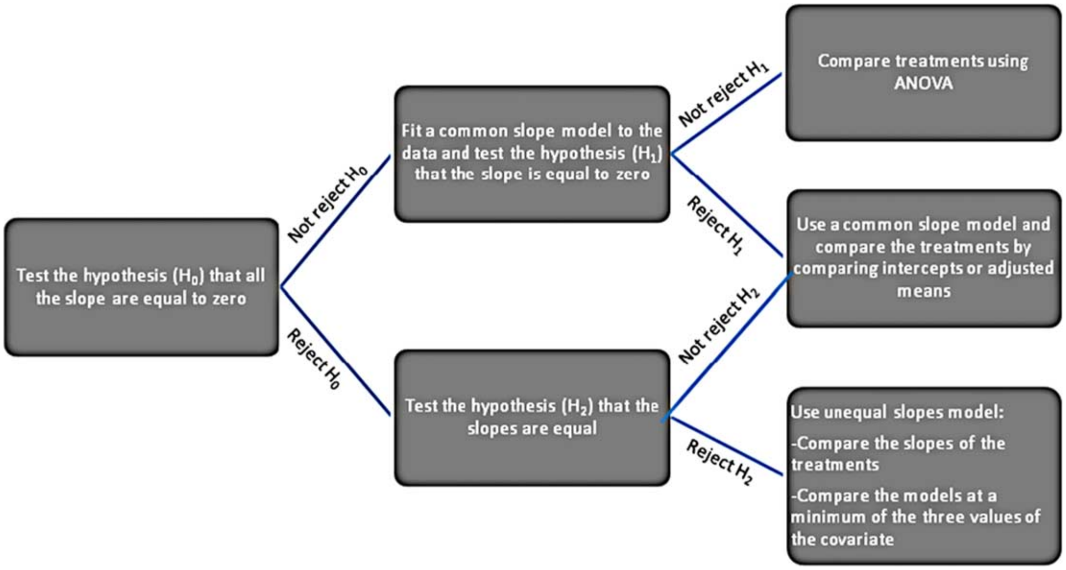
Test the hypothesis (H_0) that all the slope are equal to zero
Not reject H_0
Reject H_0
Fit a common slope model to the data and test the hypothesis (H_1) that the slope is equal to zero
Test the hypothesis (H_2) that the slopes are equal
Not reject H_1
Reject H_1
Not reject H_2
Reject H_2
Compare treatments using ANOVA
Use a common slope model and compare the treatments by comparing intercepts or adjusted means
Use unequal slopes model:
-Compare the slopes of the treatments
-Compare the models at a minimum of the three values of the covariate

Fig. 2. Flow chart for an anlysis of covariance strategy.

value from the use of ANCOVA in agricultural research, thereby stimulating more use of this powerful technique.

ACKNOWLEDGEMENTS

This paper is based on an invited talk given for the Statistical Workshop at the 2010 Canadian Society of Agronomy annual conference in Saskatoon, Saskatchewan. A sincere thank-you goes to Yantai Gan for organizing and inviting the first author to speak at the Workshop.

Brownie, C., Bowman, D. T. and Burton, J. W. 1993. Estimating spatial variation in analysis of data from yield trials: A comparison of methods. Agron. J. **85**: 1244–1253.

Cochran, W. G. 1957. Analysis of covariance: its nature and uses. Biometrics **13**: 261–281.

Cochran, W. G. and Cox, G. M. 1957. Experimental designs. 2nd ed. John Wiley, New York, NY.

Cox, D. R. and McCullagh, P. 1982. Some aspects of analysis of covariance. Biometrics **38**: 541–561.

Eberhart, S. A. and Russell, W. A. 1966. Stability parameters for comparing varieties. Crop Sci. **6**: 36–40.

Finlay, K. W. and Wilkinson, G. N. 1963. The analysis of adaptation in a plant breeding programme. Aust. J. Agric. Res. **14**: 742–754.

Fisher, R. A. 1932. Statistical methods for research workers. 4th ed. Oliver and Boyd, Edinburgh, UK.

Gomez, K. A. and Gomez, A. A. 1984. Statistical procedures for agricultural research. 2nd ed. John Wiley & Sons, New York, NY.

Lin, C. S. and Binns, M. R. 1994. Concepts and methods for analyzing regional trial data for cultivar and location selection. Plant Breed. Rev. **12**: 271–297.

Littell, R. C., Stroup, W. W. and Freund, R. J. 2002. SAS for linear models. 4th ed. SAS Institute, Inc., Cary, NC.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. 2006. SAS for mixed models. 2nd ed. SAS Institute, Inc., Cary, NC.

Milliken, G. A. and Johnson, D. E. 2002. Analysis of messy data. Volume III. Analysis of covariance. Chapman & Hall/CRC, Boca Raton, FL.

Petersen, R. G. 1994. Agricultural field experiments: Design and analysis. Marcel Dekker, Inc., New York, NY.

SAS Institute, Inc. 2008. SAS OnlineDoc 9.2. SAS Institute Inc., Cary, NC. [Online] Available: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm.

Smith, A. B., Cullis, B. R. and Thompson, R. 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J. Agric. Sci. **143**: 449–462.

Snedecor, G. W. and Cochran, W. G. 1980. Statistical methods. 7th ed. Iowa State University Press, Ames, IA.

Steel, R. G. D., Torrie, J. H. and Dickey, D. A. 1997. Principles and procedures of statistics: a biometrical approach. 3rd ed. McGraw-Hill, New York, NY.

Stroup, W. W., Baenziger, P. S. and Mulitze, D. K. 1994. Removing spatial variation from wheat yield trials: A comparison of methods. Crop Sci. **86**: 62–66.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W. and Mayo, O. 1983. Nearest neighbor (NN) analysis of field experiments. J. R. Statist. Soc. B **45**: 151–211.

Williams, E. R., John, J. A. and Whitaker, D. 2006. Construction of resolvable spatial row-column designs. Biometrics **62**: 103–108.

Yang, R.-C., Ye, T. Z., Blade, S. F. and Bandara, M. 2004. Efficiency of spatial analyses of field pea variety trials. Crop Sci. **44**: 49–55.

Yang, R.-C. 2010. Towards understanding and use of mixed-model analysis of agricultural experiments. Can. J. Plant Sci. **90**: 605–627.

Yates, F. and Cochran, W. G. 1938. The analysis of groups of experiments. J. Agric. Sci. **28**: 556–580.

APPENDIX A: A BRIEF OVERVIEW OF MIXED-MODEL THEORY FOR ANALYSIS OF COVARIANCE

GLM Model

The standard linear model, as used by the GLM procedure of the SAS system, is one of the most commonly used statistical models:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A1}$$

where **Y** is a vector of observed data, **β** is an unknown vector of fixed-effect parameters with known design matrix **X**, and **ε** is an unknown random error vector modeling the random noise around **Xβ**. The focus of GLM is to model the mean of **Y** in terms of the fixed-effect parameters **β**. The residual errors **ε** are assumed to be independent and identically distributed Gaussian random variables, each with mean 0 and variance σ_e^2.

MIXED Model

The MIXED model employed by PROC MIXED (Milliken and Johnson 2002; Littell et al. 2006) generalizes the GLM model as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{A2}$$

where **u** is an unknown vector of random-effects parameters with known design matrix Z, and **ε** is an unknown random error vector whose elements are no longer required to be independent and homogeneous as in model A1. In other words, model A2 stipulates that the vector of observations (**Y**) can be written as a sum of fixed treatment effects (**Xβ**), random block effects (**Zu**) and random experimental errors (**ε**). It is usually assumed that the random vector **u** has a multivariate normal distribution with mean vector **0** and covariance matrix **G**, i.e., **u** ~ N(**0**, **G**), and the random vector **ε** is distributed **ε** ~ N(**0**, **R**), respectively,

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right)$$

Thus, the expected value of **Y** is $\mathrm{E}(\mathbf{Y}) = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$ and the variance of **Y** is $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}) = \mathrm{Var}(\mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \mathbf{V}$ with the prime (′) representing matrix transposition. Variance-covariance matrices, **G**, **R** and thus **V** can be represented or modeled by a variety of covariance structures ranging from the simplest structure of identical and independent error variances to the most complex unstructured matrix (Littell et al. 2006). If $\mathbf{R} = \sigma_e^2\,\mathbf{I}$ and $\mathbf{Z} = \mathbf{0}$, then the MIXED model A2 reduces to the GLM model A1.

Mixed Models for Analysis of Covariance

In model A1, the elements of **X** can either represent observed values of independent variables if a regression analysis is conducted or represent dummy (0, 1) variables indicating the presence or absence of a classification effect when an ANOVA is carried out. In an ANCOVA, however, the consideration is now given to the case where some elements of **X** (called $\mathbf{X}_1$) are observed values of independent variables or covariates (called $\mathbf{X}_2$) and others are dummy (0, 1) variables or classifications. The MIXED model for ANCOVA is:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{A3}$$

where $\boldsymbol{\beta}_1$ is a vector consisting of the grand mean and the fixed effects of the classification and $\boldsymbol{\beta}_2$ is a vector of the partial regression coefficients. Following Littell et al. (2006) and Yang (2010), we obtain the best linear unbiased estimators (BLUEs) of fixed effects ($\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$) and best linear unbiased predictor (BLUP) of random effects ($\hat{\mathbf{u}}$) from solutions to the standard mixed model equations,

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\end{bmatrix}^{-} \begin{bmatrix}\mathbf{X}_1'\mathbf{R}^{-1}\mathbf{Y}\\ \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{Y}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y}\end{bmatrix} \tag{A4}$$

with superscript minus one ($^{-1}$) and superscript minus ($^{-}$) representing matrix inverse and matrix generalized inverse, respectively. The solutions in A4 assume that variance-covariance matrices, **G** and **R**, are known. In practice, **G** and **R** are usually unknown and thus, their estimates, $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$, are substituted into equation A4 to obtain empirical BLUEs (EBLUEs) of $\boldsymbol{\beta}_1$and $\boldsymbol{\beta}_2$, and empirical BLUP (EBLUP) of **u**.

We use the plant density example from Snedecor and Cochran (1980, Table 18.5.2) as analyzed in the Conventional Analysis section to illustrate how to construct appropriate matrices for obtaining mixed-model solutions in equation A4. This example is a one-way classification experiment in a randomized complete block design (RCBD) with six varieties of corn per block and a total of four blocks giving 6×4 = 24 plots. The yields *Y* (lbs/plot) and number of plants per plot (*X*) were measured. Thus, **Y** is a 24 × 1 vector of observed yields, $\mathbf{X}_1$ is a 24×7 design matrix for overall mean and six varieties, $\mathbf{X}_2$ is a 24 × 1 vector of observed numbers of plants per plot and **Z** is a 24×4 design matrix for four blocks. These matrices are given as follows:

$$\mathbf{Y}=\begin{bmatrix}202\\145\\188\\201\\202\\228\\165\\201\\185\\231\\178\\221\\191\\203\\185\\238\\198\\207\\134\\180\\220\\261\\226\\204\end{bmatrix}\mathbf{X}_1=\begin{bmatrix}1&1&0&0&0&0&0\\1&0&1&0&0&0&0\\1&0&0&1&0&0&0\\1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\\1&1&0&0&0&0&0\\1&0&1&0&0&0&0\\1&0&0&1&0&0&0\\1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\\1&1&0&0&0&0&0\\1&0&1&0&0&0&0\\1&0&0&1&0&0&0\\1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\\1&1&0&0&0&0&0\\1&0&1&0&0&0&0\\1&0&0&1&0&0&0\\1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\end{bmatrix}\mathbf{X}_2=\begin{bmatrix}28\\23\\27\\24\\30\\30\\22\\26\\24\\28\\26\\25\\27\\28\\27\\30\\26\\27\\19\\24\\28\\30\\29\\24\end{bmatrix}\mathbf{Z}=\begin{bmatrix}1&0&0&0\\1&0&0&0\\1&0&0&0\\1&0&0&0\\1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&1&0\\0&0&0&1\\0&0&0&1\\0&0&0&1\\0&0&0&1\\0&0&0&1\\0&0&0&1\end{bmatrix}$$

In this example, it is assumed $\mathbf{R}=\sigma_e^2\mathbf{I}_{24}$ and $\mathbf{G}=\sigma_b^2\mathbf{I}_4$, where σ_e^2 and σ_b^2 are error and block variances, $\mathbf{I}_{24}$ and $\mathbf{I}_4$ are identity matrices of order 24 and 4, respectively.

APPENDIX B: DATA SETS AND SAS CODE FOR ANALYSIS OF COVARIANCE

B.1. Conventional Use of ANCOVA

```
/*Stand (x, number of plants per plot) and yield (y) (lbs field weight of ear corn) of six
varieties in RCBD with four blocks (Snedecor & Cochran 1980, Table 18.5.2). Results are
present in Table 2*/

data sc;
input variety $ block x y @@;
datalines;

a      1      28      202 b  1      23      145 c  1      27      188
d      1      24      201 e  1      30      202 f  1      30      228
a      2      22      165 b  2      26      201 c  2      24      185
d      2      28      231 e  2      26      178 f  2      25      221
a      3      27      191 b  3      28      203 c  3      27      185
d      3      30      238 e  3      26      198 f  3      27      207
a      4      19      134 b  4      24      180 c  4      28      220
d      4      30      261 e  4      29      226 f  4      24      204
;
run;

/*ANCOVA–Conventional analysis. Carrying out bivariate analysis of variance;*/
```

```
proc glm data = sc manova outstat = new;
class variety block;
model x y = block variety/ss1 ss3;
lsmeans variety/stderr tdiff;
run;

proc print  data = new;
run;

*Carrying out analysis of covariance directly;

proc mixed data = sc;
class variety block;
model y = variety block x;
lsmeans variety/diff;
run;
```

B.2. Advanced Analyses with ANCOVA

B.2.1. Application #1: Analysis of dosage response

```
/*Gomez and Gomez 1984, pages 317–327. A fertilizer trial with five nitrogen rates tested on
rice yield for two seasons. Each trial has a RCBD with three replications. */

data raw;
input season $ nitrogen r1 r2 r3 @@;
n = nitrogen; *allowing nitrogen to be a classification variable and a covariate;
datalines;

dry   0     4.891  2.577  4.541 dry  60    6.009  6.625  5.672
dry   90    6.712  6.693  6.799 dry  120   6.458  6.675  6.639
dry   150   5.683  6.868  5.692 wet  0     4.999  3.503  5.356
wet   60    6.351  6.316  6.582 wet  90    6.071  5.969  5.893
wet   120   4.818  4.024  5.813 wet  150   3.436  4.047  3.740
;
run;

/*Using the ORPOL function for constructing orthogonal polynomial coefficients for five
unequally spaced nitrogen levels (0, 60, 90, 120 and 150). Results of this are presented in
Table 3.*/

proc iml;
levels = {0 60 90 120 150};
coef = orpol(levels‘);
print coef;
quit;
run;

/*Data manipulation to obtain the desired data format for subsequent ANCOVA*/

proc sort data = raw; by season nitrogen;
run;

proc transpose data = raw out = new(rename = (_name_ = rep col1 = y));
by season nitrogen n;
var r1-r3;
run;
```

```
/*Orthogonal polynomial analysis of Anaysis I in Table 4*/
proc glm data = new;
class season nitrogen rep;
model y = season rep(season) nitrogen season*nitrogen/ss1;
random rep(season)/test;

/*Make sure the sum of coefficients for each contrast are numerically zero!!
When contrasts are not significant, pool to create the remainder effect.*/

Contrast 'Linear' nitrogen −0.7278 −0.2080 0.0520 0.3119 0.5719;
Contrast 'Quadratic' nitrogen 0.4907 −0.4729 −0.4595 −0.1160 0.5577;
Contrast 'Cubic' nitrogen −0.1677 0.6312 −0.2170 −0.6213 0.3748;
Contrast 'Quartic' nitrogen 0.0368 −0.3671 0.7342 −0.5507 0.1468;
run;

/* Rerun glm to partition nitrogen*season into sub-effects.*/

proc glm data = new;
class season nitrogen rep;
model y = season rep(season) n n*n nitrogen season*n season*n*n season*nitrogen/ss1;
*random rep(season)/test;
run;

/*ANCOVA approach to dosage response
In the following SAS code, (i) Variable n is not in the CLASS statement and thus is treated as
a covariate, but nitrogen is a CLASS variable.
(ii) SS1 option is used ... Why not SS3 option??
Answer: Do not use Type III SS as adjustments of linear and linear x type interaction effects
for quadratic effects produce nonsense results!
In fact, PROC GLM or PROC MIXED do not provide F-tests for n-related effects under SS3. Now
components are broken down as shown in Analysis II of Table 4.
*/

proc mixed data = new method = type1 covtest;
class season nitrogen rep;
model y = season n n*n nitrogen season*n season*n*n season*nitrogen; */htype = 1;
random rep(season);
run;

/* To further break down error into its components */

proc mixed data = new method = type1 covtest;
class season nitrogen rep;
model y = season n n*n nitrogen season*n season*n*n season*nitrogen/htype = 1;
random rep(season) n*rep(season) n*n*rep(season);
run;

/*Plot data to see if you can find patterns in the data as shown in Figure 1. Run data to get
estimates of predicted response in the two seasons.
PROC GLM treats all effects as fixed effects when estimating these effects. Thus, the
estimates are biased.*/

proc plot data = new;
plot y*nitrogen = season;
run;

proc glm data = new;
class season nitrogen rep;
```

```
model y = season rep n(season) n*n(season)/ss1 solution;
random rep;
estimate ‘beta_0–dry season’ intercept 3 rep 1 1 1 season 3 0/divisor = 3;
estimate ‘beta_0–wet season’ intercept 3 rep 1 1 1 season 0 3/divisor = 3;
run;

/*With PROC MIXED, the RANDOM rep(season) statement and the NOINT option cause the
intercepts to be estimated directly.*/

proc mixed data = new method = type1;
class season nitrogen rep;
model y = season n(season) n*n(season) /noint solution;
random rep(season);
run;

/*PROC MIXED with METHOD = TYPE1 provides the same partitioning of SS as PROC GLM and it
gives correct F-tests for all effects that PROC GLM does not for some effects.*/

proc mixed data = new method = type1 covtest;
class season nitrogen rep;
model y = season n n*n nitrogen season*n season*n*n season*nitrogen/htype = 1;
random rep(season) n*rep(season) n*n*rep(season);
run;

quit;
```

B.2.2. Application #2: Stability of treatments across environments

```
/*Littell et al. (2002) SAS for linear models, 4th edition. Pp. 420–431.
A study was carried out to compare 3 treatments (trt) conducted at 8 locations (loc). At each
location, a RCBD design was used, but the number of blocks varied: 3 blocks inlocations 1–4,
6 blocks in locations 5 & 6, and 12 blocks in locations 7 & 8.*/

data mloc;
 input exp_unit loc blk trt y @@;
 datalines;
 1 1 1 1 46.6 2 1 2 1 43.7 3 1 3 1 37.9 4 2 1 1 34.0 5 2 2 1 38.1
 6 2 3 1 28.5 7 3 1 1 42.7 8 3 2 1 27.7 9 3 3 1 39.6 10 4 1 1 39.5
 11 4 2 1 53.4 12 4 3 1 50.2 13 5 1 1 48.0 14 5 2 1 45.8 15 5 3 1 39.0
 16 5 4 1 38.3 17 5 5 1 42.6 18 5 6 1 37.1 19 6 1 1 30.1 20 6 2 1 33.8
 21 6 3 1 35.6 22 6 4 1 33.3 23 6 5 1 31.4 24 6 6 1 39.6 25 7 1 1 34.8
 26 7 2 1 38.3 27 7 3 1 39.8 28 7 4 1 41.8 29 7 5 1 31.0 30 7 6 1 43.3
 31 7 7 1 41.1 32 7 8 1 32.9 33 7 9 1 35.0 34 7 10 1 38.0 35 7 11 1 51.5
 36 7 12 1 36.2 37 8 1 1 44.5 38 8 2 1 48.2 39 8 3 1 46.4 40 8 4 1 53.0
 41 8 5 1 51.7 42 8 6 1 43.5 43 8 7 1 44.1 44 8 8 1 43.3 45 8 9 1 44.2
 46 8 10 1 54.6 47 8 11 1 52.1 48 8 12 1 44.9 49 1 1 2 46.4 50 1 2 2 43.6
 51 1 3 2 39.5 52 2 1 2 28.5 53 2 2 2 40.0 54 2 3 2 42.5 55 3 1 2 38.9
 56 3 2 2 46.2 57 3 3 2 45.1 58 4 1 2 47.2 59 4 2 2 59.0 60 4 3 2 50.7
 61 5 1 2 46.3 62 5 2 2 53.6 63 5 3 2 44.0 64 5 4 2 41.6 65 5 5 2 44.2
 66 5 6 2 46.0 67 6 1 2 27.7 68 6 2 2 36.9 69 6 3 2 35.7 70 6 4 2 41.2
 71 6 5 2 36.5 72 6 6 2 51.0 73 7 1 2 43.6 74 7 2 2 48.9 75 7 3 2 44.6
 76 7 4 2 52.1 77 7 5 2 38.5 78 7 6 2 36.8 79 7 7 2 37.8 80 7 8 2 46.2
 81 7 9 2 43.9 82 7 10 2 41.5 83 7 11 2 48.4 84 7 12 2 47.9 85 8 1 2 48.2
 86 8 2 2 57.6 87 8 3 2 44.1 88 8 4 2 46.7 89 8 5 2 56.1 90 8 6 2 52.1
 91 8 7 2 54.8 92 8 8 2 49.4 93 8 9 2 54.6 94 8 10 2 56.6 95 8 11 2 44.3
 96 8 12 2 43.3 97 1 1 3 44.4 98 1 2 3 31.4 99 1 3 3 48.2 100 2 1 3 20.1
101 2 2 3 29.5 102 2 3 3 17.1 103 3 1 3 47.5 104 3 2 3 48.8 105 3 3 3 47.4
106 4 1 3 74.4 107 4 2 3 71.6 108 4 3 3 75.1 109 5 1 3 38.3 110 5 2 3 51.1
111 5 3 3 44.4 112 5 4 3 56.6 113 5 5 3 47.3 114 5 6 3 44.3 115 6 1 3 28.4
```

```
116 6 2 3 27.3 117 6 3 3 31.6 118 6 4 3 31.6 119 6 5 3 34.2 120 6 6 3 28.3
121 7 1 3 55.4 122 7 2 3 46.9 123 7 3 3 50.9 124 7 4 3 51.7 125 7 5 3 49.3
126 7 6 3 58.4 127 7 7 3 46.5 128 7 8 3 53.9 129 7 9 3 43.8 130 7 10 3 57.7
131 7 11 3 41.6 132 7 12 3 66.1 133 8 1 3 61.9 134 8 2 3 64.9 135 8 3 3 57.0
136 8 4 3 57.5 137 8 5 3 46.9 138 8 6 3 61.4 139 8 7 3 59.9 140 8 8 3 63.0
141 8 9 3 64.8 142 8 10 3 64.6 143 8 11 3 59.7 144 8 12 3 65.0
;
run;

/*Preliminary analysis of multi-environment trials*/;

proc mixed data = mloc method = reml covtest;
 class loc blk trt;
 model y = trt/ddfm = satterth; *without random loc blk(loc) loc*trt;
 lsmeans trt/diff;
run;

/*Use the following statements to calculate site index and analyse stability of treatments
across environments.*/

proc sort data = mloc;
 by loc; run;
proc means noprint data = mloc;
 by loc; var y;
 output out = env_indx mean = index;
run;
data all;
 merge mloc env_indx;
 by loc;
 run;

/*ANCOVA approach to stability analysis.*/

proc mixed data = all method = reml covtest;
 class loc blk trt;
 model y = trt trt*index/noint solution ddfm = satterth;
      *without location index as a covariate;
 random loc blk(loc) loc*trt;
 lsmeans trt/diff;
 contrast ‘trt at mean index’
  trt 1 −1 0 trt*index 45.2 −45.2 0,
  trt 1 0 −1 trt*index 45.2 0 −45.2;
run;

/* Output given in Table 5. */

proc mixed data = all;
 class loc blk trt;
 model y = trt trt*index/noint solution  ddfm = satterth;
 random loc blk(loc) loc*trt;
 lsmeans trt/at index = 30.9 diff; *index = 30.9 represents the poorest location;
 lsmeans trt/at means diff; *index = means  represents the average location;
 lsmeans trt/at index = 57.9 diff; *index = 57.9 represents the best location;
run;
quit;
```

B.2.3. Application #3: Analysis of spatial variability

```
/*Data taken from a field pea variety trial as described by Yang et al. (2004), Crop Sci.
```

```
44:49–55. Data is from one location.*/
data raw;
input plot_nr block entry yield @@;
datalines;

426  4  1   6419  310  3  1   6143  207  2  1   6219
121  1  1   6121  128  1  2   4934  422  4  2   5419
326  3  2   5203  215  2  2   4454  408  4  27  6482
210  2  27  5885  113  1  27  5698  323  3  27  5021
107  1  3   6751  320  3  3   5597  412  4  3   5336
224  2  3   5143  311  3  4   6277  419  4  4   6483
116  1  4   6172  211  2  4   5802  308  3  5   6791
118  1  5   4469  225  2  5   5234  423  4  5   5650
316  3  25  5497  219  2  25  5272  414  4  25  5417
117  1  25  5312  322  3  6   4438  425  4  6   5382
109  1  6   5856  203  2  6   4650  309  3  7   6684
401  4  7   4629  115  1  7   5280  201  2  7   3813
324  3  8   5536  218  2  8   5641  417  4  8   5377
105  1  8   5911  421  4  9   6157  120  1  9   6529
216  2  9   6039  305  3  9   4231  403  4  28  4174
226  2  28  5104  111  1  28  5103  313  3  28  5460
405  4  26  3971  222  2  26  5038  104  1  26  5194
312  3  26  5313  416  4  10  5742  214  2  10  5604
307  3  10  6661  114  1  10  6044  101  1  11  3735
428  4  11  5722  202  2  11  4209  328  3  11  5305
106  1  12  5607  402  4  12  4268  213  2  12  5723
315  3  12  5797  124  1  13  4721  424  4  13  5505
205  2  13  4317  317  3  13  5094  127  1  14  5035
411  4  14  6350  212  2  14  5655  301  3  14  5439
427  4  15  5474  220  2  15  5646  103  1  15  4306
303  3  15  5261  407  4  16  6227  208  2  16  6020
125  1  16  5034  314  3  16  5612  410  4  17  5304
221  2  17  5091  325  3  17  5133  110  1  17  5352
420  4  18  5778  204  2  18  5272  112  1  18  5598
302  3  18  5420  404  4  24  2460  228  2  24  3564
123  1  24  2210  327  3  24  3212  409  4  19  7493
227  2  19  5386  108  1  19  6951  304  3  19  5316
406  4  20  3464  223  2  20  4358  126  1  20  4443
319  3  20  4829  413  4  21  5538  217  2  21  5209
321  3  21  5288  119  1  21  5584  209  2  22  6083
122  1  22  5072  415  4  22  5057  318  3  22  4894
418  4  23  5609  306  3  23  4704  206  2  23  5517
102  1  23  4163
;
run;

/*Use the following SAS statements to calculate the covariate based onthe neighbouring
plot information for nearest neighbour adjustment (NNA).*/

proc glm data = raw;
       class block entry;
       model yield = block entry /ss3;
       output out = new r = res;
       *Store residuals of individual plots in the NEW data set;
  run;
proc print data = new; run;

/*Calculate the covariate based on the residuals of neighbouring plots for NNA*/

proc sort data = new; by block plot_nr;
run;

proc transpose data = new out = new2;
var res;
by block;
run;

data new3; set new2;
array rs{*} col1−col28;
array xs{*} x1−x28;
do i = 1 to 28;
if i = 1 then xs{i} = (rs{i + 1} + rs{i + 2})/2; *covariate value for the first plot of a
block;else if i = 28 then xs{i} = (rs{i − 1} + rs{i − 2})/2; *covariate value for the last
plot of a block;
else xs{i} = (rs{i − 1} + rs{i + 1})/2; *covariate value for other plots within a block;
end;
run;

proc transpose data = new3 out = new4;
var x1 − x28;
by block;
run;

data final; merge new new4(keep = res);
run;

/*RCBD analysis without NNA*/

proc mixed data = final method = reml;
class entry block;
model yield =  entry;
random block;
lsmeans entry/diff;
ods output LSMeans = lsm_rcbd;
title ‘‘RCBD analysis’’;
run;

/*RCBD analysis with NNA*/

proc mixed data = final method = reml;
class entry block;
model yield =  entry res;
random block;
lsmeans entry/diff;
ods output LSMeans = lsm_nna;
title ‘‘Nearest neighbour analysis’’;
run;

/*Comparing variety ranks without (RCBD) and with adjustment (NNA)*/

data lsm; merge lsm_rcbd(keep = entry estimate rename = (estimate = est_rcbd))
```

```
lsm_nna(keep = estimate rename = (estimate = est_nna));
run;
/* Comparison of rankings using the two approaches RCBD and NNA. Output given in Table 6. */

proc rank data = lsm out = lsm_order descending ties = low;
var est_rcbd est_nna;
ranks rest_rcbd rest_nna;
run;

proc print data = lsm_order;
run;
quit;

proc corr data = lsm_order pearson spearman;
var est_rcbd rest_rcbd;
with est_nna rest_nna;
run;
```