Spatial Mixed Model Analysis in Varietal Selection Field Trials

Tarekegn Argaw Woldemeskel[1] and Berhanu Amsalu Fenta[2]

[1]*Department of Biometrics, Ethiopian Institute of Agricultural Research (EIAR), Addis Ababa, Ethiopia,*
[2] *Ethiopian Institute of Agricultural Research (EIAR), Melkassa Agricultural Research Center, EIAR, Adama, Ethiopia; E-mail: tare.aragaw@gmail.com*

Abstract

Spatial variation is common in varietal selection field trials and is a central problem confronting a plant breeder when comparing the varieties' genetic potential. If spatial variability is not taken into account, it can strongly bias variety estimates and result in large standard errors. There have been many methods developed for accounting for spatial variation. Of these, the spatial mixed model approach proposed by Gilmour et al. (1997) has received particular attention as it simultaneously considers three types of spatial variation to be modeled: local, global, and extraneous variations. Despite the recommendations by several authors, spatial mixed model techniques are not widely used in the crop variety evaluation program as a routine data analysis platform. We present a spatial mixed model analysis using field trials from Ethiopia. Results of spatial analysis are compared to that of randomized complete block (RCB) analysis. The investigated spatial mixed models show better data fitting, resulting in a smaller error variance than that of RCB model analysis and a substantial improvement in heritability. Thus, spatial mixed models must be routinely employed in analyzing field trials to accurately and efficiently select superior varieties that contribute to agricultural productivity.

Keywords: Spatial analysis, Spatial mixed models, Field trial, Heritability, Error variance

Introduction

Agricultural varietal selection field trials are popular in plant preceding and evaluation programs, but their design and analysis continue to provide challenges for researchers. Spatial variation is common in these field trials and is a central problem confronting a plant breeder when comparing differences between varieties. Spatial variability, if not taken into consideration, can severely bias variety estimates and create huge standard errors (Girma and Niuho, 2007 and 2008; Hawinkel *et al*., 2022). The problem of spatial variability can be addressed by sound experimental design, careful trial management, and appropriate statistical analysis (Stringer *et al*. 2012). An important aim of the analysis of varietal selection field trials is to obtain good predictions for variety performance by correcting for spatial effects (Rodriguez-Alvarez *et al*., 2016).

Spatial variability can be attributed to different sources, such as inherent

variability in the experimental units and variation induced in the units through experimental processes (Cobb, 2014). The spatial variation would be more inflated when the size of the field experiment increased at the expense of plot homogeneity in the field trials, and this would definitely produce bias and inefficiency in the parameter estimation and statistical tests. Experimental designs could account for systematic variation in the plots through the process known as blocking, and it could generally increase the precision and accuracy of experiments. However, the disadvantage of blocking is that it decreases the number of degrees of freedom associated with the residual error component, so it can reduce the power of an experiment if the number of experimental units is small. Blocking also has the potential to increase the variance if it is implemented in a non-orthogonal way (Bailey, 2008). Many elements, however, remain as spatial variation that can't be handled by blocking, strongly influencing yield and other traits. It is necessary to correct for them when analyzing field trials**.**

There have been many methods developed for accounting for spatial variation, including early methods that included adjacent plot yields as covariates. Papadakis (1937) proposed a covariate adjustment procedure for neighboring plots, which was later examined by Bartlett (1938) and Williams (1952). Wilkinson *et al*. (1983) proposed a smooth trend plus independent error model with the trend being removed through differentiating the data. Removing trends by differencing neighboring plots was also used by Green *et al*. (1985) and Besag & Kempton (1986). There have been many other approaches to spatial analysis, including the one dimensional models of Gleeson & Cullis (1987), where trends were modelled using time series models, and their extension to two dimensions by Cullis & Gleeson (1991), using a separable correlation structure.

Gilmour *et al*. (1997) developed spatial models by proposing three key types of spatial variation to be modeled: local, global, and extraneous variation. Data from plots close together is more similar to those further apart, resulting in a local smooth trend that may represent changes in soil fertility, moisture, or depth. Extraneous variation is frequently caused by poor trial management or experimental procedures, for example, poor serpent planting, harvesting, or fertilizer application. To model global trend and extraneous variation, Gilmour *et al*. (1997) proposed applying polynomial or spline functions to row or column co-ordinates. The authors used a correlation structure for the residuals modeled with a separable process involving a first order autoregressive model for both rows and columns to represent local trends.

Taye and Njuho (2008) proposed a non-parametric technique for fertility

trend using P-spline smoothing. They also advocated modeling fertility trends and local variation simultaneously. To model local variation, they employed Papadakis and kriged covariates.

The aim of this paper is to present spatial analysis of individual field trials using the spatial mixed model approaches proposed by Gilmour *et al*. (1997), with a focus on three scales: global, local, and extraneous variation. The paper will also help bridge the gap between the development and application of method of spatial analysis in the analysis of individual field trials. More importantly, the paper provides the first step of analysis, called separate analysis, for individual trials enhanced with spatial analysis, which will then be combined and subjected to GxE analysis.

Material and Methods

Motivating field trials' dataset

The spatial analysis is illustrated using grain yield data from the 2019 and 2020 common bean variety trials conducted across four locations (Arsinegele, Asasa, Hawassa, and Melkassa) by the Ethiopian lowland pulses research program. Table 1 presents a summary of five trials (location by year combination). The minimum and maximum number of varieties sown across these trials was 75 and 110, respectively. The trials were designed as a randomized complete block (RCB) experiment with three replications and set out in a rectangle (row x column) array of plots, with each block being a complete duplicate. The number of common entries across trials (trial connectedness) is large (Table 2), and GxE analysis is doable, but this is not our objective in this paper.

Table 1. Summary of trials: trial dimension, number of entries, trial mean yield (t/ha), and number of missing values

Trial name	Trial dimension		Number of entries	Mean yield (t/ha)	Number of Missing values
	Column	Row			
AN19CBN1	22	22	110	2.54	0
AN20CBN1	15	15	75	2.55	0
AS19CBN1	15	9	45	6.57	6
HW20CBN1	15	6	30	1.66	0
MK20CBN1	15	15	75	2.7	0

Table 2. Common entries between trials

	AN19CBN1	AN20CBN1	AS19CBN1	HW20CBN1	MK20CBN1
AN19CBN1	110				
AN20CBN1	69	75			
AS19CBN1	45	28	45		
HW20CBN1	28	30	28	30	
MK20CBN1	69	75	28	30	75

Spatial mixed model

Consider a grain yield dataset collected from a varietal selection field trial in which *m* varieties are grown. The trial consists of n_{plots} arranged in a rectangular array with *c* columns by *r* rows ($n = c$ x r). Let $\boldsymbol{y}_{be}$ the (n x 1) data vector for a trial, ordered as rows within columns

The linear mixed model for y can be then written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{\alpha}$ is vector of fixed effects (including terms for the grand mean, global spatial trend, and other fixed effects) with an associated design matrix $\mathbf{X}$ (assumed to be full column rank), $\mathbf{u}$ is the a vector of random effects (including terms for variety, replication, extraneous spatial variation, and other random effects) with associated design matrix $\mathbf{Z}$, and $\boldsymbol{\varepsilon}$ is the (*n x 1)* vector of residual errors ordered as for the data vector.

The random effects from the linear mixed model (equation 1) are assumed to follow a Normal distribution with mean zero vector and variance-covariance matrix, that is

$$\sigma^2 \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix},$$

Where $\mathbf{G}$ is variance matrix for random effects and $\mathbf{R}$ is the variance-covariance matrix of the residual error

Individual trial residual effects can be analyzed employing spatial methods of analysis that account for local or plot-to-plot variation. $\mathbf{R}$ in this case will have its own spatial covariance structure. Spatial mixed model provides for a much more realistic correlation structure between plots. In the linear mixed model, we partition the (*n x 1*) vector of residual errors $\boldsymbol{\varepsilon}$ into a vector $\boldsymbol{\psi}$ (n × 1) of spatially correlated effects, and a vector $\boldsymbol{\zeta}$ of independent technical or measurement errors. Thus, for variety field trials that have row by column arrangement and ordered as rows within columns, the spatially correlated effects can be modelled using a two dimensional separable correlation structure (Gilmour *et al.*, 1997). Thus, the covariance matrix for $\boldsymbol{\psi}$ is given by

$$\text{var}(\boldsymbol{\psi}) = \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Sigma}_c(\boldsymbol{\rho}_c) \otimes \boldsymbol{\Sigma}_r(\boldsymbol{\rho}_r) \qquad 2$$

Where $\boldsymbol{\Sigma}_{\mathbf{c}}$ and $\boldsymbol{\Sigma}_r$ represent spatial correlation structures with parameters in $\boldsymbol{\rho}_c$ and $\boldsymbol{\rho}_r$ for the column and row directions, respectively. In both the row and column directions, we typically use an autoregressive spatial structure of order one, with in $\boldsymbol{\rho}_c$ and $\boldsymbol{\rho}_r$ each containing a single autocorrelation parameter.

For spatial auto-correlation in the row direction only, the model simplifies to $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{n_c} \otimes \boldsymbol{\Sigma}_r(\boldsymbol{\rho}_r)$ where n_c is the number of columns for the trial. Similarly, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Sigma}_c(\boldsymbol{\rho}_c) \otimes \mathbf{I}_{n_r}$ would

be the reduced form for spatial auto-correlation in the column direction only where n_r is the number of rows for the trial. We can have also no spatial covariance in either direction. Thus, the model simplified to an IID variance structure of the form $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_{n_c} \otimes \mathbf{I}_{n_r} = \sigma^2 \mathbf{I}_n$.

Heritability formula

Following the approach of Cullis *et al.* (2006), the heritability (H^2) value for the trial can be calculated from a generalized formula that takes unbalanced data into account, as

$$H^2 = 1 - \frac{A}{2\sigma_g^2} \quad (3)$$

Where A is the average pairwise prediction error variance of genetic effects is σ_g^2 is the genetic variance.

Estimation, testing and software

Estimation in the linear mixed model involves estimating the fixed and random effects, **α** and **u** and the variance-covariance parameters in **G** and **R**. This involves two linked processes, in which the variance parameters of the model are estimated using residual maximum likelihood (REML, Patterson & Thompson 1971), and the fixed and random effects are estimated using best linear unbiased estimation (BLUE) and best linear unbiased prediction (BLUP) respectively. The Residual Maximum Likelihood Ratio Test (REMLRT) can be used to test the significance of random effects in the linear mixed model. The REMLRT can only be used to compare the fit of two models that are nested and have the same fixed effects.

The Wald test may be used to determine the significance of fixed effects in a linear mixed model. The classic Wald statistic has an asymptotically chi-squared distribution. This test is often seen as anti-conservative (Butler *et al*., 2009). Kenward and Roger (1997) introduced an adjusted Wald statistic along with a F approximation that worked well in a range of scenarios. ASReml in the R environment was used to estimate the variance parameters from the linear mixed model using REML (Butler *et al*., 2009). The Average Information (AI) algorithm is implemented by ASReml (Gilmour *et al*., 1995).

Results and Discussion

The first model fitted was a randomized complete block (RCB) model with random block/replication and variety effects, and the residual correlation structure is denoted by id(Column).id(Row), where id refers to the identity matrix. This model reflects the RCB analysis based on mixed model approach. And, the next step was to model the spatial variation. Thus, a spatial model to the residuals using a separable autoregressive process in the column and row dimensions was first fitted over the RCB model for the local variation. The spatial model for extraneous variation was then fitted along the

column and row by retaining significance terms for local variation. The significance of each fitted spatial model for both local and extraneous variation was tested using the REMLR test. And finally, the spatial models for global variation were fitted and tested with the Wald test.

The presence of global and extraneous variation associated with the row and column direction can be checked with the help of model diagnosis tool, in particular the sample variogram (Gilmour *et al*. 1997). A sample variogram as a typical model diagnostic plot for trial AN20CBN1 using the Metplot package, which is a part of ASReml-R, is depicted in Figure 1. In the row direction, the plot has a saw-toothed up-down pattern. This trend, as demonstrated by Stefanova *et al*. (2009), is most likely the presence of extraneous variation within the trial, implying random row effects. By including a random column effect in the mixed model, this effect can be accommodated in the model. There is also evidence of a global trend in the row direction, with the variogram steadily increasing. This global trend may be accounted for in the model by employing linear or polynomial regression over a centered row number, denoted lrow.

Table 3 presents the Wald test for global variation and the REMLR test for local and extraneous spatial variation. The spatial variation were appropriately modeled for AN19CBN1 along the row and column (p-value <0.001), for HW20CBN1 along the row (p-value=0.017), and for MK20CBN1 along the column(p-value=0.01). The global and extraneous spatial variation were found to be significant along the row direction at AN20CBN1 (p-value =0.019 and <0.001, respectively). This demonstrates how spatial variation is a crucial component of plant breeding trials that must be considered. The estimates of the local spatial correlation parameters were positive ($\rho1$=(0.33, 0.57), $\rho2$ =0.38, $\rho3$=0.18 , see Table 3), suggesting what was noted in De Faveri (2013) that plots near together are more likely to be comparable than plots further apart. Negative spatial correlation values would imply inter-plot competition, as shown in Stringer and Cullis (2002), but there was no indication of it here.

Control non-genetic variability by blocking, which is based on observable factors like soil type and topography, is commonly used in designing experiments, but several possible inputs are unknown or unmeasured, which stresses the importance of incorporating spatial information into the analysis of each trial to improve the precision of variety evaluation. As noted in Girma (2005), spatial analysis is critical for field trials in Ethiopia, where the variability of soil and environment is very high, even within a very small area. If this spatial variation is unaccounted for, it leads to biased and inefficient results for variety valuation.

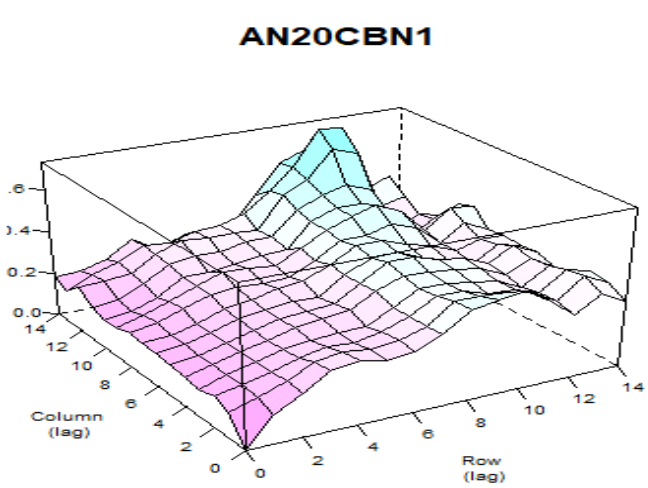

Figure 1. Sample variogram plot of the residuals from a base model for yield at trial MK18N.

Table 3. Summary of spatial analysis: spatial variation, fitted model term, Wald and REML test statistic and P-value

Trial name	Spatial variation	Model terms	Wald[a] /REMLR[b] test statistics	P-value
		ar1(Column):id(Row)	47.59	<0.001
AN19CBN1	Local	ar1(Column):ar1(Row)	65.27	<0.001
	Global	lrow	7.18	0.019
AN20CBN1	Extraneous	Row	16.74	<0.001
AS19CBN1	Extraneous	Column	4.17	0.041
HW20CBN1	Local	id(Column):ar1(Row)	5.68	0.017
MK20CBN1	Local	ar1(Column):id(Row)	6.58	0.01

ρ1=(0.33, 0.57), ρ2 =0.38, ρ3=0.18

[a]Test for global trend after significant terms for extraneous variation and local trend are fitted
[b]Test for extraneous variation and local trend after significant terms for global trend are fitted
ρ1, ρ2 and ρ3 are estimates for autoregressive order 1(AR1) spatial correlations parameters at, AN19CBN1, HW20CBN1 and MK20CBN1, respectively

The genetic variance, error variance and heritability estimates for each trial from both methods of analysis (RCB and spatial analysis) are presented in Table 4. All results are REML estimates, and are unbiased for the corresponding variance component parameters. In the RCB analysis the estimates ranged from 0.044 to 0.297 for genetic variance, from 0.076 to 5.686 for error variance, from 29.49 to 63.33 for heritability. Similarly, in the spatial analysis the estimates ranged from 0.044 to 1.215 for genetic variance, from 0.08 to 4.246 for error variance, and from 42.82 to 76.92 for heritability. At AS19CBN1, both analyses revealed very high genetic and error variance, as well as low heritability. This might be because the spatial variation captured in the spatial analysis was considerable, and the remaining variability of the experimental units (plots) was also extremely large. The RCB analysis gives low genetic variance for most of the trials. This is not

surprising since the RCB analysis is quite inferior to the spatial analysis to efficiently explaining the genetic variance and spatial variations in the data.

As compared to RCB analysis, the spatial analysis gave relatively small error variance for all trails. It substantially also increased heritability in all trials as it clearly depicted in Figure 2. This in general reveals that the spatial analysis will improve the precision and accuracy of varieties evolution by capturing non-genetic variation associated with agricultural field plot variability (Smith and Cullis, 2018; Cullis *et al*. 2010).

Table 4. Trial genetic variance, error variance and heritability from RCB and spatial analysis

	RCB analysis			Spatial analysis		
Trial name	Genetic variance	Error variance	Heritability	Genetic variance	Error variance	Heritability
AN19CBN1	0.297	0.564	61.21	0.297	0.555	76.92
AN20CBN1	0.103	0.339	47.71	0.125	0.175	66.94
AS19CBN1	0.835	5.686	29.49	1.215	4.246	42.82
HW20CBN1	0.044	0.076	63.33	0.044	0.08	67.55
MK20CBN1	0.059	0.188	48.6	0.057	0.08	48.68

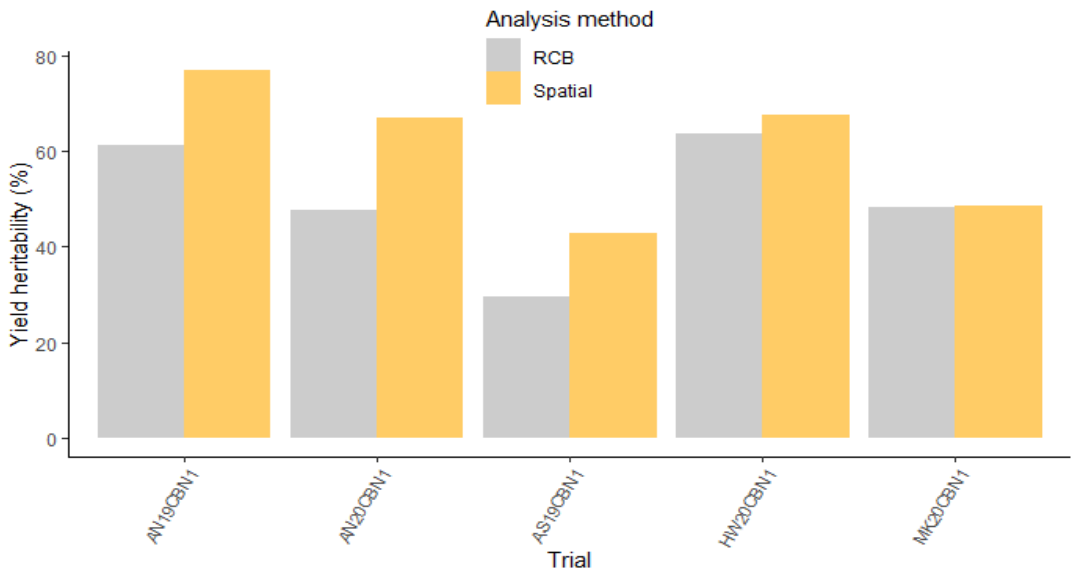

Figure 2. Heritability of yield at each trial using RCB and spatial analysis.

Conclusion

Agricultural field trial data (such as variety trials) may not be balanced or may be big trials at the expense of plot homogeneity, and ANOVA-based approaches may not be appropriate for their analysis. Spatial variation is

clearly evident in varietal trials. If this spatial variation is not taken into account, variety selection will worsen. As a result, more efficient spatial analysis approaches must be used. The linear mixed model provides a strong framework for dealing with unbalanced as well as relaxing the distributional assumptions surrounding the residual error by employing spatial analysis. The investigated spatial models show better data fitting, resulting in a smaller error variance than that of RCB model analysis and a substantial improvement in heritability. For each individual variety trial, spatial variation on three scales: global, extraneous, and local trend must be considered to obtain accurate and efficient results for variety evaluation.

Acknowledgements

Dr. Alison Kelly for some helpful spatial analysis suggestions, Common bean breeders for organizing the trials row data, and an unnamed referees for extremely helpful remarks that enhanced the presentation of the work are all acknowledged by the authors.

References

Bailey, R.A., 2008. *Design of comparative experiments* (Vol. 25). Cambridge University Press.

Bartlett, M.S., 1938. The approximate recovery of information from replicated field experiments with large blocks. *The Journal of Agricultural Science*, *28*(3), pp.418-427.

Besag, J. and Kempton, R., 1986. Statistical analysis of field experiments using neighbouring plots. *Biometrics*, pp.231-251.

Butler, D.G., Cullis, B.R., Gilmour, A.R. and Gogel, B.J., 2009. ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.

Butler, D.G., Cullis, B.R., Gilmour, A.R. and Gogel, B.J., 2009. ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.

Cobb, G., 2014. Introduction to design and analysis of experiments. John Wiley, New York.

Cullis, B.R. and Gleeson, A.C., 1991. Spatial analysis of field experiments-an extension to two dimensions. *Biometrics*, pp.1449-1460.

De Faveri, J., 2013. *Spatial and temporal modelling for perennial crop variety selection trials* (Doctoral dissertation).

Gilmour, A.R., Cullis, B.R. and Verbyla, A.P., 1997. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, pp.269-293.

Gilmour, A.R., Cullis, B.R. and Verbyla, A.P., 1997. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, pp.269-293.

Gleeson, A.C. and Cullis, B.R., 1987. Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics*, pp.277-287.

Green, P., Jennison, C. and Seheult, A., 1985. Analysis of field experiments by least squares smoothing. *Journal of*

the Royal Statistical Society: Series B (Methodological), *47*(2), pp.299-315.

Hawinkel, S., De Meyer, S. and Maere, S., 2022. Spatial regression models for field trials: a comparative study and new ideas. *Frontiers in plant science*, p.592.

Papadakis, J.S., 1937. Bull. Inst. Amél. Plantes à Salonique..

Patterson, H.D. and Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), pp.545-554.

Rodríguez-Álvarez, M.X., Boer, M.P., van Eeuwijk, F.A. and Eilers, P.H., 2016. Spatial models for field trials. *arXiv preprint arXiv:1607.08255*.

Smith, A., Cullis, B. and Gilmour, A., 2001. Applications: the analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics*, *43*(2), pp.129-145.

Smith, A.B. and Cullis, B.R., 2018. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, *214*(8), pp.1-19.

Stefanova, K.T., Smith, A.B. and Cullis, B.R., 2009. Enhanced diagnostics for the spatial analysis of field trials. *Journal of agricultural, biological, and environmental statistics*, *14*(4), pp.392-410.

Stringer, J.K. and Cullis, B.R., 2002. Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early stage sugarcane selection trials. *Australian Journal of Agricultural Research*, *53*(8), pp.911-918.

Stringer, J.K., Smith, A.B., Cullis, B.R. and Hinkelmann, K., 2012. Spatial analysis of agricultural field experiments. *Design and Analysis of Experiments, Volume 3: Special Designs and Applications*, *810*, p.109..

Taye, G. and Njuho, P., 2007. An improvement on the papadakis covariate to account for spatial variation. *Journal of agricultural, biological, and environmental statistics*, *12*(3), pp.397-413.

Taye, G. and Njuho, P.M., 2008. Smoothing fertility trends in agricultural field experiments. *Statistics*, *42*(3), pp.275-289.

TAYE, G., 2005. *Using spatial modelling techniques to improve data analysis from agricultural field trials* (Doctoral dissertation, University of KwaZulu-Natal Republic of South Africa).

Ullis, B.R.C., Mith, A.B.S. and Oombes, N.E.C., 2006. On the Design of Early Generation Variety. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(4), pp.381-393.

Wilkinson, G.N., Eckert, S.R. and Hancock, T.W., 1983. TW and Mayo, O. *Nearest neighbour (NN) analysis of field experiments (with discussion). JR Statist. Soc. B*, *45*, pp.151-211.

Williams, R.M., 1952. Experimental designs for serially correlated observations. *Biometrika*, *39*(1/2), pp.151-167.