International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Proper Data Analysis Techniques to Reduce Experimental Error in Longitudinal Data

W.H.H.Fernando[a]*, K. P.Waidyarathne[b], D.D.M.Jayasundara[c]

[a,c]*Department of Statistics & Computer Science, University of Kelaniya, Kelaniya 11600, Sri Lanka*

[b]*Plant Physiology Division, Coconut Research Institute, Lunuwila, Lunuwila 61150, Sri Lanka*

[a]*Email: himashaharshani635@gmail.com*

[b]*Email: Pramuditha.waidyarathne@gmail.com*

[c]*Email: jayasund@kln.ac.lk*

Abstract

This study presents findings of research conducted to improve analysis techniques of experimental data from coconut research. It highlights the ways of handling unaccountable variability due to the inconsistent temporal behavior of the experimental units in perennial crop research to obtain a precise research output. Properly designed field experiments are essential to identify the influence of independent variable/s on the dependent/s at the various stages. This document highlights the ways how the improved methodologies can be successfully used to reduce the experimental error in most commonly used experimental designs and types of analysis. The first example, the study on long term coconut fertilizer experiment designed as a randomized complete block design in Badalgama, Sri Lanka, compares different types of analyses via evaluating the model residuals and calculating the coefficient of variability (CV) to reduce the error and thereby improve the output. The statistical methods used in the first case study includes Repeated Measure Analysis of Variance (RMANOVA) as the classical method and Repeated Measure ANOVA (With single palm per plots), Linear Mix Model, and MANOVA with two Principal Components that represent approximately 80% variation of the data as dependent variables as improved methods. The model adequacy of each approach was accepted after testing normality, homogeneity of variance and independence of residuals. CV resulted from classical RMANOVA was 39.95%, while it was 39.2% from Repeated Measure ANOVA (With single palm per plots) and 16.51% from the Linear Mixed model.

----

* Corresponding author.

The lowermost CV (10.04%) resulted from MANOVA with two principal components indicating that it can be more powerfully used to analyze long term experiments of coconuts. The second example, the study on long term coconut fertilizer experiment from Bandirippuwa, Sri Lanka that failed to have normality assumption of parametric methods, illustrates appropriate types of the Non-Parametric analysis(F2-LD-F1) for the longitudinal data. The regularity of the results should be studied further with few more comparable data sets.

Keywords: Coefficient of Variations; Non-Parametric analysis; Principal Components; Randomized Complete Block Design; Repeated Measure Analysis of Variance.

1. Introduction

Designing, conducting, and analyzing field experiments are crucial to the success of any research program. In perennial crop research, different types of experiments are mainly conducted to evaluate fertilizer and agronomic trials and to screen breeding trials. These experiments are widely conducted as randomized complete block designs. Blocks in the fields are laid out perpendicular to the observed gradients and in such a way that plots within each block are as uniform as possible before the application of treatments. Treatments are randomly allocated in plots within each block. The classical methods are used to analyze the data from such experiments. However, it has been long term observed that these methods do not take into account the heterogeneity of experimental units, which inflate the experimental error by different temporal behaviors, thus end results are low in precision. This will eventually result in ambiguous conclusions from the experiments. The tall form of Coconut, the commercially grown coconut cultivar in Sri Lanka, is a unique heterozygous genotype. Therefore, a special emphasis should be given to design field research with coconut palms due to multiple years of data collection and existing high variation between individuals. Due to the importance of experimental designs in coconut research, early researchers have suggested several methods to improve the power of the experimental designs over five decades [1]. Past studies have not taken much effort to develop new field designs but to select proper plot sizes for field experiments by comparing the reduction of variability and thereby the error. The optimum plot size was initially concluded as 16-18 palms[2], then reduced up to 6 palms[3] and more recently as a single palm [4]. Even though the single palm plots are recommended in the literature, practical uses of such designs are hardly found. To date, the most common experimental design used in coconut research is Randomized complete block design (RCBD) with 6 effective palms in a plot. However, coconut research still suffers from high experimental error due to heterogeneity of coconut palms, their different performances in weather variations and finer variability that cannot be visually recognized in experimental fields may mask true treatment effects. In coconut field experiments, the same palms are being measured over several years on the same dependent variable (eg: Yield) mainly due to the perennial nature of the crop and to account for the temporal dynamics of the experimental patterns. Repeated measures analysis of variance (RMANOVA) is, therefore, used in coconut experimentation as the classical data analysis method to detect any overall differences between related means. RMANOVA has an advantage over independent ANOVA as this has the effect of increasing the value of the F-statistic due to further partitioning of within subject variability (error in ANOVA) into variability in subjects and error leading to an increase in the power of the test. The error of RMANOVA reflects individual variability to each time (how subjects react to different conditions/time). However, heterogeneity and unpredictable behavior of coconut palms in different years more often cause violation of the sphericity assumption making within subject RMANOVA statistics are meaningless. This is due to the temporal fluctuation within plot variability due to different environmental responses of palms regardless of the treatment imposed. High variation among the individuals in similarly treated plots makes treatment mean sensitive to those fluctuations ultimately masking the true treatment effect. Even careful planning of the experiment cannot ensure total elimination of this component. Therefore, there should be a proper methodology to handle this uncertainty and thereby improve coconut experimentation. Therefore, there is a real need of addressing the current issues of failure of experimental designs used in coconut research and improve them for better experimental planning. This study is based on an analysis of secondary data from existing long-term field experiments of coconut to optimize data analyzing techniques for experimental designs in coconut research. The main purpose of the study is to enhance the precision of data interpretation of coconut field experiments by improving the method of experimental data analysis. The value of the relative efficiency of each modified technique will be evaluated to understand the most suitable technique for long term data analysis in coconut experimentation. The research will demonstrate how some relatively simple computations and analytical methods can be used to improve the data analysis and thereby to improve the precision by effectively handling the noise. The main objective of this study is to improve the analyzing method of repeated measures data from coconut experiments with an improved methodology.

2. Materials and Methods

2.1 Data

Data from field experiments conducted by the coconut research institute of Sri Lanka was used in the study. These experiments were purposefully selected as treatments were not significantly different according to the results of conventional analysis. The response variable in the experiment was coconut yield harvested at bi-yearly intervals. The first experiment was initiated in 2006 in Badalgama area for determining the effect of five different fertilizer combinations and continued for 8 years (from 2006 to 2013). The experiment was designed as a completely randomized block design with four replicates and five treatments. Each similarly treated plot contained six palms. The second experiment was initiated in 2006 in Bandirippuwa Estate for determining the effect of four different fertilizer combinations and continued for 12 months (from 2006 January to December). The experiment was designed as a completely randomized block design with three replicates and four treatments. Each similarly treated plot contained six palms.

2.2 Data Analysis Methods

2.2.1 Classical Analysis of Data: Repeated Measure ANOVA

Coconut experiments have the practice of measuring the outcome on each coconut tree multiple times. Most of the time, the outcome is the yield in bi-monthly intervals for several years, which have repeated exposure to changing levels of weather. Repeated measures analyses are often conducted with RMANOVA. RMANOVA is a member of the ANOVA family, which is used to compare group means on a dependent variable across repeated measurements of time. RMANOVA model includes zero or more independent variables and at least 1 dependent variable that has more than one observation as shown in equation 1.

$$Y_{i,j,k} = \mu + \tau_i + \alpha_j + \beta_k + \varepsilon_{i,j,k} \quad (1)$$

Where, $Y_{i,j,k}$ is the observation/measurement taken from i^{th} treatment in j^{th} block at k^{th} time point, τ_i is the effect of i^{th} treatment, α_j is the effect of j^{th} block, β_k is the parameter associated with the k^{th} time and $\varepsilon_{i,j,k}$ is the error term. RMANOVA is similar to the dependent sample T-Test, because it also compares the dependent means of the same individuals over different times. It is necessary for the RMANOVA for the observations in one-time point to be directly linked with the observations in all other time points. This automatically happens when repeated measures are taken, or when analyzing similar units or comparable specimen. When adapting the ANOVA for repeated measurements, it is essential to assume a common set of periods or schedule among all the individual units. This requirement can be easily met in agricultural studies. Time is often referred to as the within-subjects factor, whereas a fixed or no changing variable (e.g., treatment) is referred to as the between-subjects factor[5]. The rationale underlying the RMANOVA analysis is to consider time as a factor on k levels in a hierarchical design with individual units (or subjects) as subplots. RMANOVA must meet the assumption of normality of residuals by time point and sphericity, sometimes referred to as compound symmetry. Sphericity requires that the repeated measures demonstrate homogeneity of variance and homogeneity of covariance. Homogeneity of covariance means that the relationships, or correlations, on the dependent variable among all of the repeated measures are equal. The probability of type I error increases, if the data set does not satisfy the sphericity assumption. In such cases, Greenhouse-Geisser (GG) and Huynh-Feldt (HF) methods can be considered as correction procedures to modify the degrees of freedom values of the time and the error/residual to accomplish the correction. However, the application of GG adjustment assumes a maximum violation of sphericity assumption hence use of GG for correcting minimal violations may enhance the type II error (accepting false H_0). Therefore, the literature suggests [6] to apply the HF correction when epsilon is > 0.75 and GG correction when epsilon is < 0.75. Specification of the within subject variance covariance structure is the key step in the analysis of repeated measures. There are several variance covariance structures available for selection by analysts, and many statistical software procedures implement the compound symmetry structure by default. The compound symmetry variance covariance structure assumes that all the variances are equal and all the covariances are equal. These assumptions, however, might be incorrect and a different structure might better describe the variance between subjects and covariation within subjects. Repeated measures designs, however, have some disadvantages compared to designs that have independent groups. Dependency in the data created by the repeated measures should be taken into the account when analyzing repeated measures designs. In reality, the scores of the dependent variable are not independent on each other and the dependency may bias the results because each participant responds to several stimuli, making the responses more similar within participant than across participants. This, in turn, creates a correlation (dependency) among repeated measures, and this correlation should be incorporated in the statistical models to avoid (or minimize) biases. In a design with response times as dependent variable, for instance, one can observe that participants tend to be different in their average speed of response, independently to the experimental conditions. If so, some participant will be always slower or faster than others, creating a correlation among measures. One way to capture dependency due to repetition within participants and repetition within stimuli is to employ a mixed model in which the model coefficients are allowed to vary across participants and across stimuli. That is, a model with random coefficients.

2.2.2 Principle Components to Adjust Dependent Variables

Considering each of these two procedures, multivariate analysis; PCA (principal-components analysis), are largely used as dimension-reducing procedures for a collection of continuous variables. These techniques can identify a small set of synthetic variables, called eigenvectors or factors.[7] In this study, the longitudinal observations are dependent over time. We have 8 dependent variables in the first experimental data set that use to PCA on for the purpose of data reduction and the result of the violation of the sphericity in the Repeated Measure ANOVA method. Essentially, these measures represent eight different times of coconut yield. After running the PCA analysis, check the components with eigenvalues over 1 which cumulatively account of the variance. The K-M-O measuring sampling adequacy gives an overall measure of variance. Then the K-M-O test values can be justified for PCA.[8] By using those principle components, construct the Multivariate ANOVA (MANOVA)[9]. Due to the independence of the chosen principal components, we can use the MANOVA.

2.2.3 Linear Mixed Model ANOVA

The general linear mixed model approach has both fixed and random effects in experiments as shown in equation 2.

$$y = X\beta + Zu + \varepsilon \tag{2}$$

Where y is a N×1 column vector, the outcome variable; X is a N×p matrix of the p predictor variables; β is a p×1 column vector of the fixed-effects regression coefficients (the βs); Z is the N×qJ design matrix for the q random effects and J groups; u is a qJ×1 vector of q random effects (the random complement to the fixed β) for J groups; and ε is a N×1 column vector of the residuals. It also has the potential of accommodating multiple missing data points[10] and higher order, nonlinear changes in the dependent measure across time. With these broad possibilities for modeling longitudinal data, the mixed model approach is becoming immensely popular in the experimental literature. The mixed model has several unique abilities to;

- Characterize group and individual behavior patterns in a formal way
- Acknowledge both group and individual differences
- Incorporate additional covariates

The mixed model is a more subject specific model and a natural choice for analyzing longitudinal data as it naturally represents individual trajectories in a formal way. Unlike the RMANOVA, which requires a complete balanced array of data, the mixed model can accommodate a dataset with a large portion missing. Although the RMANOVA requires a fixed schedule among all individual units, the mixed model can accommodate flexible time schedules. Furthermore, rather than treating time as a categorical variable, as in the RMANOVA, the mixed model is capable of treating time as either a continuous variable or a categorical variable or both. Adaptation of time as a continuous variable allows for varied entry of participants into a study and that also allows for several, generally nonequivalent possibilities for modeling behavior, for example, mixed-effects, marginal, and transitional models[11].

2.2.4 Nonparametric Analysis of Repeated Measures Data

In most real-world situations, the distribution of observed data is unidentified and there may exist a number of distinctive measurements and outliers. Consequently, in practice, specific model assumptions of parametric procedures can rarely be verified. If dependent variables do not satisfy the model assumptions (normality, the randomness of the error), parametric statistical procedures may result in unreliable conclusions. Therefore, the use of parametric and semiparametric techniques that impose restrictive distributional assumptions on observed longitudinal samples becomes questionable. As an alternative, nonparametric rank-based methods that can offer a flexible and robust framework for the analysis could be of help. Literature provides some evidence for using nonparametric marginal models in different types of longitudinal[12] developed R based software package for analyzing longitudinal data from commonly used factorial designs. Authors introduced a notational system for each design to be denoted by Fx-LD-Fy, where x and y are the number of whole-plot and sub-plot factors, respectively, while “LD" stands for “longitudinal data". Whole-plot factors (between subjects) are the factors, which stratifies samples in independent groups (eg: Treatments), while the factors, stratifying repeated measurements, are called sub-plot factors (within-subjects). For such designs, the statistical model can be described by independent random vectors;

$X_{ijk} = (X_{ijk1}, X_{ijkt})^T$; k = 1, nij; with marginal distributions $X_{ijks} \sim F_{ijs}$; i = 1..., a; j = 1…., b; and s = 1…, t. The total number of observations is N = n.t, where $n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$.

The nonparametric hypotheses of no main effect (A), no main time effect (T), and no interaction (AT) between A and T, are expressed in terms of the marginal distribution functions:

$$H_0^F(A): \overline{F_{1.}} = \cdots = \overline{F_{a.}}$$

$$H_0^F(T): \overline{F_{.1}} = \cdots = \overline{F_{.t}}$$

$$H_0^F(AT): F_{is} = \overline{F}_{i.} - \overline{F}_{.s} + \overline{F}_{..},$$ i= 1…. a; s = 1…t,

where $\overline{F}_{i.} = \frac{1}{t}\sum_{s=1}^{t} F_{is}$ denotes the mean distribution over time for treatment group i; i =

1,…….a, $\overline{F_{.s}} = \frac{1}{a}\sum_{i=1}^{a} F_{is}$ denotes the mean distribution over the treatment groups for time point s; s = 1,…… t, and $\overline{F}_{..} = \frac{1}{at}\sum_{i=1}^{a}\sum_{s=1}^{t} F_{is}$ denotes the overall mean distribution.

2.3 Efficiency Calculations

2.3.1 Coefficient of Variation

Coefficient of variation (CV) is a measure of relative variability (eq: 3). It is the ratio of the standard deviation to the mean (average). CV is particularly useful when it is required to compare results from two different surveys or tests that have different measures or values.

The formula for the CV is:

$$CV = \left(\frac{Standard\ Deviation}{Mean}\right) * 100\% \quad (3)$$

When comparing different analysis methods, in particularly ANOVA, mean square error (MSE) of ANOVA can be used as the standard deviation of the design because standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation.

3. Results & Discussion

Of the two data sets used in the study, data set one satisfied the assumptions required for parametric data analysis while the assumptions were not met in the second data set even after several data transformations. Therefore, in this section, classical method (RMANOVA) of analysis of experiment design for experiment one data set vs improved methodologies (RMANOVA with single palm plots, MANOVA with two principal components, Linear Mixed Model) were compared to study how efficient those methods are to reduce the experimental error and thereby increase the precision of experimental design. Next, the section explains how nonparametric methods were used to analyze experimental data when the normality assumption of the classical method (Repeated Measure ANOVA) was violated using the second data set.

3.1Classical Repeated Measure ANOVA (Experiment -one Analysis)

Most statistical techniques assume certain characteristics of the data. A valid interpretation of results requires that one or more such assumptions be satisfied. Therefore, the assumptions of RMANOVA were checked as a violation of these assumptions changes the conclusion of the research and interpretation of the results.

3.1.1Normality

RMANOVA assumes that the test variables follow a multivariate normal distribution in the population. Therefore, normality of the data set was checked with the Shapiro-Wilk test (Table 1) as it is more appropriate for small sample sizes (< 50 samples). As raw data did not satisfy the normality assumption, data were square root transformed to achieve the normality. All the test variables with square root transformation satisfied the normality assumption as they were not significant at the 5% level of the significance (Table 1).

Table 1: Tests of normality for annual per palm yield data of experiment 1 from 2006 to 2013 according to Shapiro-Wilk (SW) test

Year	Raw Data		Square Root Transformed	
	Statistic	Sig.	Statistic	Sig.
2007	0.97	0.01	0.99	0.73
2008	0.96	0.01	0.99	0.75
2009	0.99	0.37	0.99	0.60
2010	0.99	0.33	0.99	0.69
2011	0.98	0.17	0.98	0.26
2012	0.98	0.13	0.98	0.28
2013	0.94	0.00	0.98	0.27

3.1.2 Sphericity

The most important assumption in RMANOVA is that variances of all differences between all possible pairs of repeated measures factor must be equal in the population. The violation of sphericity should be seriously considered as it causes an increase in the type I error rate making within subject RMANOVA statistics are meaningless. Sphericity was tested with Mauchly's test and results were given in table 2.

Table 2: Tests of sphericity for annual per palm yield data of experiment 1 from 2006 to 2013 according to Mauchly's test

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon	
					Greenhouse - Geisser	Huynh-Feldt
Time	0.05	288.59	35	0.00	0.49	0.55

According to table 2, the sphericity assumption was violated indicating that the variance of coconut palms in different years were not equal. Therefore, degrees of freedom for the averaged tests of significance were adjusted using Greenhouse Geisser correction as the epsilon estimate was below 0.75. Adjusting df for lack of sphericity using Greenhouse Geisser correction, within subjects' effects (Time effect) were compared and results revealed that the time effect was significant (df=3.75, F=9.82 and sig.=0.00). This indicates that there are significant differences between repeated measures. However, results further revealed that Time * Treatment interaction was insignificant (sig. = 0.06) indicating that temporal behaviour of each treatment was consistent at 95% confidence. However, with 1% increase of the probability level, Time * Treatment interaction becomes significant giving an indication that temporal behaviour of different treatments can be different due to high heterogenic behaviour among individual palms.

Table 3: Tests of between-subjects' effects with Greenhouse Geisser correction in classical RMANOVA

Source	Sum of Squares (Type III)	df	Mean Square	F	Sig.
Intercept	1,390.00	1	1,390.00	172.06	0.00
Cov.	527.10	1	527.10	65.25	0.00
Treatment	18.32	4	4.58	0.57	0.69
Block	49.21	3	16.40	2.03	0.11
Error	815.91	101	8.08		

Tests of between subjects in classical RMANOVA (Table 3) indicated that treatments were not significantly different. Results indicated that insignificance of treatment effect may either be due to insignificant treatment effect or low precision of analysis methods due to high variability between subjects. In calculating the efficiency of RMANOVA in terms of CV, the mean square error (8.08) of the model was used as the standard deviation of the design. Residual analysis of the model confirmed that residuals were independent, normally distributed having a constant error variance (Table 4).

Table 4: Tests of normality with Shapiro-Wilk test, independence with Box-Pierce test and constant error variance with Levene Test

Residual for Year	Box-Pierce test		Shapiro-Wilk		Levene Test	
	X-squared	Sig.	Statistic	Sig.	Statistic	Sig.
2007	1.08	0.30	0.994	0.89	3.16	0.02
2008	0.07	0.79	0.989	0.51	2.26	0.07
2009	1.80	0.18	0.988	0.46	0.54	0.71
2010	3.69	0.05	0.989	0.53	0.20	0.94
2011	0.01	0.91	0.978	0.07	0.65	0.63
2012	2.09	0.15	0.988	0.44	0.35	0.85
2013	1.81	0.18	0.992	0.77	1.08	0.37

3.2 Repeated Measure ANOVA (With single palm plots)

In the previous analysis, we considered five treatments in four blocks in which each plot contained six palms. Therefore, in classical RMANOVA, an average of six palms was taken to represent the effect of a particular treatment. The danger of taking an average of several individuals (six palms here) to represent the plot value for a particular treatment effect is that the mean value is sensitive for extreme behaviors of palms, which is very common in coconut. Use of covariate (same variable recorded prior applying the treatments) to adjust the initial variability between treatments may have a little effect in classical analysis if individuals are highly heterogeneous. This is because, use of the mean of the covariate to adjust the plot means may not represent the true variability between individuals. And use of the previous year yield may not be that influential to control the variability if the temporal behavior is not consistent among individuals. However, we assumed that if we consider single palm plots for the analysis, initial variability can be adjusted with the prior measurements of the same palm without averaging. Therefore, RMANOVA with a single palm plot approach was used to analyze the data with the aim of reducing the experimental error. As there were six palms in each plot and a block contains five such plots (five treatments), hypothetically 7776 (6^5) single palm plot combinations can be organized in one block. Besides, there are three such blocks in the experiment. Therefore, if we are to compare the error reduction in single palm plots, we have to perform RMANOVA for a large number of combinations. In this study, we simulated 100 such designs and the results of one analysis are shown below as an example. Here, results suggested that the degrees of freedom of the within subjects' effect should be adjusted with either Greenhouse-Geisser or Huynh-Feldt correction as the sphericity assumption was violated as per the Mauchly's test (table 5).

Table 5: Tests of sphericity for annual per palm yield data of experiment 1 from 2006 to 2013 according to Mauchly's test

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Time	0.05	236.70	35	0.00	0.49	0.68	0.12

After adjustment of the df to correct the sphericity violation, the results of within subjects' effect of the experiment indicated that there is a significant time effect but all the treatments behaved similar in different times. Significance of time effect was obvious in both above analysis with very high probability of significance (sig. 0.00). According to the results of between-subjects' effects in single palm plot RMANOVA (Table 6), the MSE was 7.78 (in this particular example), which is slightly lower than what obtained in the classical RMANOVA (8.08). We computed 100 MSEs by running RMANOVA for different hypothetical combinations of single palm plots and compared them with the MSE of classical RMANOVA using single sample t test. Results revealed that MSE of single palm plot analysis was significantly lower from the classical MSE (7.82 ± 0.06, $p=0.00$) but the treatment effect remains non-significant in all the analyses.

Table 6: Results of between-subjects' effects in the single palm plot RMANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1029.48	1	1029.48	155.46	0.00
Cov	273.30	1	273.30	35.13	0.00
Block	219.41	21	10.45	1.34	0.17
Treatment	18.49	4	4.62	0.59	0.67
Error	645.72	83	7.78		

Use of single palm plots with considerable number of replicates (blocks) may help in reducing experimental error by minimizing soil variability within a block and correctly adjusting initial variability of experimental units as explained above. However, hypothetical blocking in this analysis may not help in reducing soil variability as expected in field research.

Results of the normality with Shapiro-Wilk test, independence with Box-Pierce test and constant error variance with Levene test (Table 7) proved that the assumptions of the error were not violated most of the time, hence, the analysis conducted was accepted as appropriate.

Table 7: Tests of normality with Shapiro-Wilk test, independence with Box-Pierce test and constant error variance with Levene Test

Residual for Year	Box-Pierce test		Shapiro-Wilk		Levene Test	
	X-squared	Sig.	Statistic	Sig.	Statistic	Sig.
2007	1.77	0.18	0.99	0.80	2.76	0.03
2008	0.19	0.66	0.99	0.90	3.06	0.02
2009	1.49	0.22	0.99	0.82	0.92	0.45
2010	1.18	0.28	0.99	0.51	0.51	0.73
2011	0.02	0.90	0.98	0.07	0.68	0.61
2012	3.98	0.06	0.97	0.02	0.96	0.43
2013	3.04	0.08	0.99	0.89	1.29	0.28

3.3 MANOVA with Two Principal Components as Dependent Variables

The third approach we used in the study was the formulation of a few principal components (PCs) instead of eight annual yield variables and use them as the dependent variables in the analysis. Here, we believed that projection of response variables onto a subspace using a data compression method is advantageous to reduce the variability between experimental units.

Table 8: Component Matrix of Principal Component Analysis after Varimax Rotation and Kaiser Normalization

Year	Component	
	1	2
2011	0.88	
2009	0.87	
2010	0.86	
2008	0.83	
2007	0.80	
2012		0.95
2013		0.94

Results of the principal component analysis revealed that the total variability of experimental data can be explained by two major PCs, where the first PC explained 56.27% variability, while the second explained 22.86%. Table 8 illustrates the coefficients of each year in two principle components after Varimax Rotation and Kaiser normalization. Accordingly, repeated measures (years) were grouped into two clear groups defined by the highest loading on each year, which was evident on the component plot in rotated space that showed how closely related the years were to each other and to the two components.

Then, we conducted MANOVA taking these two PCs as response variables. Results of the multivariate analysis of variance of the two PC's are shown in table 9.

Table 9: Tests of Between-Subjects Effects in MANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	42.52	1	42.52	82.74	0.00
Cov	52.66	1	52.66	102.47	0.00
Block	2.78	3	0.93	1.80	0.15
Treatment	1.68	4	0.42	0.82	0.57
Error	51.90	101	0.51		

Results revealed that the MSE was drastically reduced (0.51) compared to two previous analyses with raw data but still the treatment effect was not significant. The model adequacy was accepted after checking normality (Shapiro-Wilk – 0.99, p=0.62), homogeneity of variance and independence of residuals (Box-Pierce =5.05, p=0.25). However, even after reducing the variability between repeated measures, treatment effect was not significant in neither of components.

3.4 Linear Mixed Model Analysis

Liner Mixed model analysis was conducted as the next method of improving the analysis of coconut experiments. Here, we considered the effect of individual palms as a random effect. Results of the linear mixed model analysis are shown in table 10.

Table 10: ANOVA of the mixed model with six palms per plot design

Parameter	Value	Std Error	df	t-value	P value
Intercept	6.60	0.36	770	18.47	0.00
Block	0.10	0.10	107	0.99	0.32
Treatment	0.09	0.08	107	1.16	0.25

The model found out that the extracted standard error of the model was 1.38. Model adequacy was confirmed with residual analysis where normality of the residual analysis of the fixed effects (Shapiro-Wilk = 0.99695, p-value = 0.09) and the random effects (Shapiro-Wilk = 0.99358, p-value = 0.89) was achieved. The constant variance of the residuals was confirmed as the residuals were randomly scattered. Independence of the residuals were confirmed using Box-Pierce test (X-squared = 3.7869, df = 1, p-value =0.0522). Here also, there was no significant difference between treatments, but the probability of significance was comparatively lower than what obtained for all other analysis.

3.5 Comparison of the efficiency of the models

The calculated CVs for error terms of each method was shown in the Table 11. In both classical RMANOVA and RMANOVA (With single palm plots) resulted comparatively higher CV close to 39 - 40%, which was in agreement with the CV values reported in literature for coconut experiments [4]. However, CV was considerably reduced in linear mixed model analysis as a result of introducing palm to palm variation as a random effect into the model. MANOVA with PCs showed the lowest CV but it may be due to the use of dependent variables in a reduced dimension.

Table 11: Coefficient of variation of each analysis method

Analysis Method	Coefficient of Variation(CV)	$CV = \left(\frac{Standard\ Deviation}{Mean}\right) \times 100\%$
Classical RMANOVA	39.95%	$\sigma/\mu \times 100\% = (\sqrt{8.08}/7.11542) \times 100\% = 39.95\%$
Repeated Measure ANOVA (With single palm plots)	39.20%	$\sigma/\mu \times 100\% = (\sqrt{7.78}/7.11542) \times 100\% = 39.20\%$
ANOVA with a Principal Component as Dependent Variable	10.04%	$\sigma/\mu \times 100\% = (\sqrt{0.51}/7.11542) \times 100\% = 10.04\%$
Linear Mixed Model Analysis	16.51%	$\sigma/\mu \times 100\% = (\sqrt{1.38}/7.11542) \times 100\% = 16.51\%$

3.6 Nonparametric Analysis of Repeated Measures Data (Experiment- two Analysis)

In the second experiment, data did not follow the normal distribution to apply RMANOVA even after several data transformations. As mentioned in the methodology section, this data set was analyzed using a nonparametric method, which was an inbuilt R software model package (F2- LD- F1 model)[12]. Authors introduced a notational system for each design to be denoted by Fx-LD-Fy, where x and y are the number of whole-plot and sub-plot factors, respectively, while “LD" stands for “longitudinal data". Whole-plot factors (between subjects) are the factors, which stratifies samples into independent groups. In this data set they were treatments and blocks (x=2). The factors stratifying repeated measurements, which are called sub-plot factors (within-subjects), in the current data set was Time factor (Y=1). Accordingly, the model was in F2 –LD-F1 structure. Results of F2-LD-F1 model analysis are shown in table 12.

Table 12: ANOVA table of F2-LD-F1

	Statistic	df	p-value
Treatment	0.99	2.90	0.39
Block	1.86	1.91	0.15
Time	21.24	4.89	0.00
Treatment*Block	1.11	5.36	0.35
Treatment*Time	0.97	12.43	0.46
Block*Time	0.62	8.93	0.77
Treatment*Block*Time	0.91	20.26	0.56

The model found out that the treatments are non-significance and the time factor is significant at 5% level of significance.

4. Conclusion

RMANOVA with six palms per plot, the classical method, resulted comparatively high experimental error (coefficient of variation of error term 39.95%). The first improved method, RMANOVA with single palm plot resulted CV =39. 2%, which was nearly equal to what obtained with classical analysis. It was observed that the reduction of error was negligible. Multivariate ANOVA conducted with two main principal components reduced CV by three times (10.04%) compared to the classical method. Results clearly showed a reduction of experimental error than the classical RMANOVA too. This may be because the model used modified dependent variables in a reduced space. This method was preferable over others as each principal component was a linear combination of the original features while preserving as much as possible from the total variance of the data. So, by performing dimensionality reduction using PCA and coupling it with MANOVA, it was expected to reduce the within class variability and increases between-classes variability. The Liner mixed model, with its broad possibilities for modeling longitudinal data, also resulted comparatively lower CV (16. 51%) than classical RMANOVA but higher than Multivariate ANOVA with two PCs. None of the analyses resulted significant the treatment effect even if the error was reduced. Considering the efficiency, MANOVA with two Principal Components as dependent variables can be recommended as the better way of handling the heterogeneity problem of coconut palms but results need to be validated with more longitudinal data sets. For the type of data that violates parametric model assumptions, a non-parametric method such as F2–LD- F1would be a good method to analyze longitudinal data. However, the results cannot be compared to the results of the parametric methods.

5. Limitations & Recommendations

The biggest limitation in coconut experimentation is the high variability between experimental units and their different temporal behaviour. As being a perennial plant, it is compulsory to have multiple years data collection on the same parameters. Therefore, a better data analysing method is required to handle longitudinal data with high variability. Above analysis were experimented with the aim of reducing error is one such data set but results need to be generalized with several similar kind of data sets. In addition, literature provide evidence for applying Bayesian modelling successfully on such data. Therefore, Bayesian methodology will also be tested and compared on coconut data as it is expected to generate more accurate results with it than classical inference methods because Bayesian models have the capacity for analyzing intra-individual variability as a predictor.

Acknowledgments

I would like to express my sincere gratitude to Dr. Pramudhitha Waidyaratne Senior Biometrician, Coconut Research Institute, Lunuwila, for her advice, teaching and support have been fundamental for the elaboration of this study. Then my gratitude is conveyed to, Dr. D. D. M. Jayasudara, Senior Lecturer, Department of Statistic & Computer Science, Faculty of Science, University of Kelaniya her patience, motivation, and immense knowledge. And also grateful to W. D. D. Fernando who supported me thought this venture was greatly appreciated.

References

[1]. T. S. G. PERIES and R. O. Thattil, "ASSESSMENT OE THE EFFECTS OF ENVIRONMENTAL FACTORS ON YIELD OF COCONUT (Cocos nucifera, L.)," in Cocos, 2010, vol. 12.

[2]. W. V. D. Peiris and M. L. M. Salgado, "Experimental error in field experiments with coconuts," 1937.

[3]. V. Abeywardena, "Studies on biennial bearing tendency in coconut I. the measurement of bienniality in coconut," 1962.

[4]. J. Kularatne, T. S. G. Peiris, and S. Samita, "Feasibility of single palm plot in coconut experimentation," Trop. Agric. Res., vol. 18, p. 306, 2006.

[5]. S. W. Huck and B. H. Layne, "Checking for proportional n's in factorial ANOVA's," Educ. Psychol. Meas., vol. 34, no. 2, pp. 281–287, 1974.

[6]. N. Haverkamp and A. Beauducel, "Violation of the Sphericity Assumption and Its Effect on Type-I Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM)," vol. 8, no. October, pp. 1–12, 2017.

[7]. G. H. Dunteman, Principal components analysis, no. 69. Sage, 1989.

[8]. M. R. Abdullah, R. M. Musa, A. Maliki, N. A. Kosni, and M. A. Aziz, "The Application of Principle Components Analysis to Identify Essential Performance Parameters in Outfield Soccer Players," Res. J. Appl. Sci., vol. 11, no. 11, pp. 1199–1205, 2016.

[9]. S. Leigh, "A user's guide to principal components." Taylor & Francis, 1993.

[10]. G. G. Moisen, D. R. Cutler, and T. C. Edwards Jr, “Generalized linear mixed models for analyzing error in a satellite-based vegetation map of Utah,” Quantifying Spat. Uncertain. Nat. Resour. Theory Appl. GIS Remote Sens., pp. 37–43, 2000.

[11]. C. Krueger and L. Tian, “A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points,” Biol. Res. Nurs., vol. 6, no. 2, pp. 151–157, 2004.

[12]. K. Noguchi, Y. R. Gel, E. Brunner, and F. Konietschke, “nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments,” vol. 50, no. 12, 2012.