J. Hortic. Sci.
Vol. 18(2) : 475-479, 2023

https://doi.org/10.24154/jhs.v18i2.2172

Original Research Paper

# An alternate statistical method for dealing outliers in perennial crop experiment

**Venugopalan R.*, Kurian R.M., Chaithra M. and Sisira P.**
ICAR-Indian Institute of Horticultural Research, Bengaluru - 560089, Karnataka, India
*Corresponding author Email: venugopalan.r@icar.gov.in

## ABSTRACT

A statistical method based on Robust ANOVA to handle outliers induced high coefficient of variation (CV) in pooled (2011-2018) analysis of long-term Mango cv. Totapuri rootstock trail was suggested. Based on the results, it was concluded that the rootstock treatment T3: Olour (average yield over the period 2011 to 2018 as 57.21 kg/tree) as the best. Precision gained as estimated by reduction in CV (%) was in the range of 11.01 % to 78.9 %. SAS IML codes were built-in for analysis. Hence, this study calls for employing robust ANOVA approach in testing the significance of evaluated treatments in a designed perennial crop experiment with high CV that would have reduced the sensitivity of testing the significance of treatment differences otherwise.

**Keywords:** Mango, Outliers, Robust ANOVA

## INTRODUCTION

Classical analysis of variance (ANOVA) approach to compare the significance of set of treatments in a perennial crop field experiment is mainly based on the requirement of certain assumptions for the ANOVA model. The major hindrance to this is the presence of outlier(s) among the replicated values. Outliers in any of the replications (of any treatment) lead to failure of normality assumption. Presence of such aberrant values may finally leads to non-significance/on par results coupled with high coefficient of variation (CV), especially in perennial fruit crops spaced very widely in the open field such as mango.

One way out is to identify such an outlier(s) and delete them to have a possible comparison among treatments. However, deleting the outlying replication is not recommended because its deletion leads to violation of basic principle designs of experiment (i.e. randomization) and from experimenter point of view every observation carries some information that should be exploited. This aspect is very much pertinent especially when we deal with perennial trees, as the number of replicated values for a treatment kept at a bare minimum. To address this problem, a method based on Robust ANOVA is suggested and its efficacy is studied using primary data on yield related traits with a view to identify best treatment. Robust ANOVA techniques are designed to be less sensitive to outliers, which are data points that deviate significantly from the rest of the data. It improves the reliability of the analysis by reducing the influence of outliers (Paul & Bhar, 2011; Venugopalan & Manjunath, 2019).

## MATERIALS AND METHODS

Eight root stocks treatments such as T1: Totapuri, T2: Vellaikulumban, T3: Olour, T4: Peach, T5: Kensington, T6: Mylepelian, T7: Nekkare and T8: Turpentine were selected for the study and evaluated in RCBD, with three replications, at an experimental plot of Division of Fruit Crops, ICAR-Indian Institute of Horticultural Research, Bengaluru during the period 2010-2018 was considered. Primary data recorded on three important characters of Mango cv. Totapuri that were showing high CV of more than 20% almost consistently throughout the experiment period viz., fruit yield per tree in kilograms, average weight of individual fruit and number of fruits harvested per tree, while all other characters studied were showing less than 18% CV consistently, for eight rootstocks treatments. Both classical and Robust ANOVA were employed to identify best rootstock treatment for each of the traits.

*a) Classical two-way analysis of variance :*

The two-way ANOVA (Federer, 1975) model that describes the response variable with treatment and block effect is given by

$$Y = \mu + \alpha_i + \beta_j + \varepsilon \text{...............}(1)$$

where $\alpha_i$= effect of ith treatment, $\beta_j$=effect of jth block, $\varepsilon$=random error

When the experiment is conducted over seasons or years or places pooled ANOVA or combined analysis of data is done after the analysis of individual experiments. Before going for the pooled analysis the data is tested for homogeneity of error variance using Bartlett's Chi-square test. If the chi-square test result is significant, we go for pooled ANOVA.

**Bartlett's Chi-square test**

Null hypothesis of Bartlett chi-square test H0: The variances in the different groups are equal against the alternate hypothesis H1: The variances in the different groups un- equal, indicating heterogeneity of variances.

A. When $p$ =2

$$\chi^2 = \frac{s_{e_1}^2}{s_{e_2}^2} \sim \mathrm{F}_{(n1,\,n2)} \text{ ..........................(2)}$$

Where $S_{e_1}^2$ and $S_{e_2}^2$ are mean square error for two years or seasons or places.

B. When $p$ >2

$$\chi_{p-1}^2 = \frac{\sum n_i \log S_e^{-2} - \sum n_i \,\mathrm{lo} \quad {}_{e_i}^{2}}{1+\frac{1}{3(p-1)}(\sum\frac{1}{n}-\frac{1}{\sum n_i})} \sim \chi_{p-1}^2, \text{ ..................(3)}$$

where $S_e^{-2} = \frac{\sum n_i S_{e_i}^2}{\sum n_i}$, $S_{e_i}^2$ is the mean sum of square of ith year with $n_i$ df

When the calculated value is more than the critical value the test is significant, which means null hypothesis is accepted and pooled ANOVA is performed. If the chi – square result is not significant, indicating the heterogeneity of error variance and to stabilize the error variance, appropriate transformation is chosen, then pooled ANOVA could be performed.

*b) Robust Analysis of variance :*

Robust M-estimation approach instead of minimizing the sum of squared residuals in the classical ANOVA based approach, minimizes the sum of a less rapidly increasing function of the residuals ( $\rho\left(e_i\right)$), as given below (Paul & Bhar,2011; Venugopalan & Manjunath, 2019).

$$Min\sum_{i=1}^{n}\rho\left(y_i - \sum x_{ij}\beta_j\right) = Min\sum_{i=1}^{n}\rho\left(e_i\right) \text{...............(4)}$$

The solution is not scale equi-variant, and thus the residuals must be standardized by a robust estimate of their scale $\overset{\wedge}{\sigma}_e$ which is estimated simultaneously. As in the case of M-estimates of location, the median absolute deviation (MAD) is often used. Taking the derivative of above equation and solving, produces the score function

$$\sum_{i=1}^{n}\Psi\left(y_i - \sum x_{ij}\beta_j / \overset{\wedge}{\sigma}\right)x_{ik} = \sum_{i=1}^{n}\Psi\left(e_i / \overset{\wedge}{\sigma}_e\right)x_i = 0 \text{ .........(5)}$$

Where $\overset{\wedge}{\sigma} = median\left|e_i - median\left(e_i\right)\right| / 0.6745$

With $\Psi = \rho'$. There is now a system of k+1 equations, for which $\psi$ is replaced by appropriate weights that decrease as the size of the residual increases

$$\sum_{i=1}^{n} w_i\left(e_i / \overset{\wedge}{\sigma}_e\right)x_i =$$

$$\sum_{i=1}^{n} \frac{x_{ij}\{\psi[(y_i - x_i'\beta)/s]/(y_i - x_i'\beta)/s\}/(y_i - x_i'\beta)}{s} = 0 \text{ .....(6)}$$

$\longrightarrow$ j=0.1,.....,k

As

$$\sum_{i=1}^{n} x_{ij}w_{i0}\left(y_i - x_i'\beta\right) = 0 \text{ ..............(7)}$$

$\longrightarrow$ j=0.1,.....,k

where

$$w_{i0} = \begin{cases} \dfrac{\psi\left[\left(y_i - x_i'\overset{\wedge}{\beta}_0\right)/s\right]}{\left(y_i - x_i'\overset{\wedge}{\beta}_0\right)/s} & \text{if } \mathrm{y_i} \neq x_i'\overset{\wedge}{\beta}_0 \\ 1 & \text{if } \mathrm{y_i} = x_i'\overset{\wedge}{\beta}_0 \end{cases}$$

Hence by matrix notation $\mathrm{X'W_0}X\beta = X'\mathrm{W_0}y$

where $\mathrm{W_0}$ is n × n diagonal matrix of weights then one step estimator is -

$$\overset{\wedge}{\beta} = \left(\mathrm{X'W_0}X\right)^{-1}X'\mathrm{W_0}y \text{ ...............(7)}$$

**Robust criterion functions**

| Criterion | $\rho(z)$ | $\psi(z)$ | $w(z)$ | Range |
|---|---|---|---|---|
| Least squares | $\frac{1}{2}z^2$ | z | 1.0 | $\lvert z\rvert < \infty$ |
| Huber's function | $\frac{1}{2}z^2$<br>$\lvert z\rvert t - 1/2t^2$ | z<br>t sign(z) | 1.0<br>$t/\lvert z\rvert$ | $\lvert z\rvert \le t$<br>$\lvert z\rvert > t$ |

Here $\rho(z)$ is the function of residual, $\psi(z)$ is the derivative of and w (z) is the weight function (Huber, 1973). SAS codes using SAS V 9.3 were generated for both the estimation procedures and used for analysis (SAS V 9.3, 2012).

***Comparison of classical ANOVA vs Robust ANOVA***

Efficacy of set of treatments in both the approaches are tested by computing the p-value (a measure of strength of the inference drawn) and the coefficient of variation (Gomez and Gomez,1988). The results are presented in Tables 2-4.

## RESULTS AND DISCUSSION

For the entire study we took eight root stock treatments for 3 characters such as yield / tree, fruit weight, number of trees during 2011-2017 respectively. The results of both classical and robust ANOVA methods for three characters studied are presented in Tables 2-4. Individual year based assessment of significant treatments as revealed by the respective P-values of the treatment along with CV values are computed and presented.

It may be observed that for the character Yield/tree, except for 2017-18, all the treatments are *on par* to each other (Table 2), as the respective p-value exceeded 0.05. However, for the trait average fruit weight, all the individual year based classical ANOVA resulted in *on par* results (Table 3),as the respective p-value exceeded 0.05. Further, for the trait number of fruits, except for 2017-18, all the treatments are on par to each other (Table 4), as the respective p-value exceeded 0.05. However, in most of the analysis, the value of coefficient of variation exceeded 20%, a cut of value desired for any field based experimental study.

Since, there is significance of results among the treatments in some of the years of experiment, before going for the pooled analysis, the data is tested for homogeneity of error variance using Bartlett's Chi-square test. In our study the preliminary results showed heterogeneity in error variance, hence transformed the original values using logarithmic transformation and the proceeded for pooled ANOVA. Perusal of the results presented in Table 1 justified for proceeding to pooled analysis of variance as the computed $x^2$ values, for the all the three traits supported for the presence of heterogeneous error variance. Accordingly, the results of pooled ANOVA for all the three traits are presented in the last row of first two columns of Table 2-4. Similar trends as observed in individual year based analysis was also observed in pooled ANOVA based results, leading to inability for identifying the best rootstock treatment. This may probably due to the presence of outliers in one or two replications across treatments in some of the individual year based analysis. Outliers are values that are unusually far from the main concentration of data points. These extreme observations can skew and mislead the statistical analysis, leading to inaccurate conclusions if not properly addressed or accounted for. Accordingly, robust ANOVA method was employed and the results are presented in Table 2-4.

Perusal of the results of robust ANOVA for the trait yield / tree revealed a significant difference among all the treatments tested during all the years and also for the pooled data, since the probability value being less than 0.05. There is a considerable reduction in the value of coefficient of variation in most of the cases to lesser than 20%, with the pooled data resulting in CV value as 14.16%. The precision gained (as computed as the reduction in CV due robust ANOVA over classical ANOVA) due to the robust ANOVA approach for individual year based and pooled analysis is presented in the penultimate column of Table 2. It was observed that the precision gained by this approach was high as 11.53%. The remain other two traits (Table 3 and 4) with the precision gained being around 60.37 and 34.97% respectively. A DMRT based post-hoc test was adopted to suggest the best treatment for all the three traits individually. The results presented in the last column of the respective tables revealed that the Olour rootstock for Totapuri scion (T3) as the best for both the traits, yield / tree and average fruit weight, however the rootstock treatment Turpentine for Totapuri (T8) is the best for number of fruits, and was on par with T3.

**Table 1 : Results of Bartlett's test for individual traits**

| **Character** | $\chi^2$ Cal |
|---|---|
| Yield / tree | 2.77** |
| Average fruit weight | 8.61** |
| Number of fruits | 4.23** |

** significance at p<0.05

**Table 2 : Comparison of regular ANOVA and Robust ANOVA methods for yield/tree**

| Year | Classical Pooled ANOVA | | | Robust Pooled ANOVA | | | Reduction in CV (%) precision gained | Best treatment (as per robust pooled) |
|---|---|---|---|---|---|---|---|---|
| | P-value Treatment | P-value Treat vs. Year | CV (%) | P-value Treatment | P-value Treat vs. Year | CV (%) | | |
| 2011 | 0.77 | NA | 34.59 | 0.77 | NA | 34.60 | 11.54 | - |
| 2012 | 0.04 | NA | 16.79 | 0.03 | NA | 16.75 | 8.22 | - |
| 2013 | 0.40 | NA | 14.06 | 0.34 | NA | 14.09 | 7.47 | - |
| 2014 | 0.57 | NA | 10.51 | 0.57 | NA | 10.46 | 0.46 | - |
| 2015 | 0.41 | NA | 11.21 | 0.408 | NA | 11.23 | 10.71 | - |
| 2016 | 0.43 | NA | 8.65 | 0.432 | NA | 8.71 | 7.75 | - |
| 2017 | 0.006 | NA | 9.58 | 0.0008 | NA | 10.57 | 7.09 | - |
| Pooled | 0.001 | 0.682 | 44.70 | 0.001 | 0.68 | 14.160 | 68.32 | $T_3$ |

**Table 3 : Comparison of regular ANOVA and Robust ANOVA methods for average fruit weight**

| Year | Classical Pooled ANOVA | | | Robust Pooled ANOVA | | | Reduction in CV (%) precision gained | Best treatment (as per robust pooled) |
|---|---|---|---|---|---|---|---|---|
| | P-value Treatment | P-value Treat vs. Year | CV (%) | P-value Treatment | P-value Treat vs. Year | CV (%) | | |
| 2011 | 0.22 | NA | 2.82 | 0.218 | NA | 2.75 | 2.54 | - |
| 2012 | 0.36 | NA | 1.26 | 0.364 | NA | 1.49 | 13.34 | - |
| 2013 | 0.123 | NA | 1.24 | 0.126 | NA | 0.49 | 60.37 | - |
| 2014 | 0.14 | NA | 1.30 | 0.139 | NA | 1.32 | 21.75 | - |
| 2015 | 0.17 | NA | 1.26 | 0.174 | NA | 1.12 | 11.22 | - |
| 2016 | 0.81 | NA | 1.26 | 0.808 | NA | 1.40 | 4.23 | - |
| 2017 | 0.53 | NA | 1.86 | 0.49 | NA | 1.15 | 37.86 | - |
| Pooled | 0.178 | 0.213 | 1.79 | 0.18 | 0.21 | 1.68 | 5.81 | $T_3$ |

**Table 4 : Comparison of regular ANOVA and Robust ANOVA methods for number of fruits**

| Year | Classical Pooled ANOVA | | | Robust Pooled ANOVA | | | Reduction in CV (%) Precision gained | Best treatment (as per robust pooled) |
|---|---|---|---|---|---|---|---|---|
| | P-value Treatment | P-value Treat vs. Year | CV (%) | P-value Treatment | P-value Treat vs. Year | CV (%) | | |
| 2011 | 0.836 | NA | 15.52 | 0.84 | NA | 25.01 | 34.97 | - |
| 2012 | 0.026 | NA | 12.95 | 0.026 | NA | 12.81 | 1.10 | - |
| 2013 | 0.369 | NA | 9.87 | 0.34 | NA | 11.17 | 10.88 | - |
| 2014 | 0.746 | NA | 7.00 | 0.75 | NA | 7.56 | 21.96 | - |
| 2015 | 0.279 | NA | 9.10 | 0.28 | NA | 8.58 | 5.72 | - |
| 2016 | 0.477 | NA | 6.10 | 0.43 | NA | 8.71 | 1.79 | - |
| 2017 | 0.530 | NA | 6.58 | 0.510 | NA | 11.50 | 16.38 | - |
| Pooled | 0.001 | 0.613 | 11.07 | 0.001 | 0.613 | 11.09 | 0.72 | $T_8$ |

## ACKNOWLEDGEMENT

Authors are grateful to the Director, ICAR-IIHR, for providing all facilities to carryout the work.

## REFERENCES

Gomez, K.A., Gomez, K.A., & Gomez, A.A. (1984). Statistical procedures for agricultural research. John Wiley & Sons.

Huber, P.J. (1973). Robust regression: Asymptotic, conjectures, and Monte carlo. *Annals of Statstics, 1*, 799-821.

Federer, W.T. (1955). Experimental design: theory and application. Macmillan, New York.

Paul, R.K., & Bhar, L.M. (2011). M-estimation in block design. *Journal of Indian Society of Agricultural Statistics, 65*(3), 323-330.

SAS V 9.3 2012. Statistical analysis system version 9.3 SAS Institute, Cary NC.

Venugopalan, R., & Manjunath, B.L. (2019). Application of Robust ANOVA methods in Papaya having outlier data. *Journal of the Indian Society of Agricultural Statistics, 73*(2), 129-132.

***(Received : 23.08.2023; Revised : 11.10.2023; Accepted 27.10.2023)***